Genomewide Analysis of Proteins that Bind to DNA and Regulate Gene Expression, With Particular Emphasis on Imprinted Genes

Hughes, Siobhan Mary

Awarding institution:
King's College London

Genomewide Analysis of Proteins that Bind to DNA and Regulate Gene Expression, With Particular Emphasis on Imprinted Genes

Submitted by

Siobhan Mary Hughes
November 2016

To King's College London for the degree of
Doctor of Philosophy

The work submitted in this thesis is my own.

Abstract

Regulation of gene expression is a complicated process, subject to different mechanisms operating at different levels. At the genomewide level, work in this thesis describes the use of chromatin immunoprecipitation and next-generation sequencing to interrogate the coincident and allele-specific binding of the proteins CTCF, cohesin, ATRX and MeCP2. Identifying regions co-binding these proteins helps generate a model to understand how these proteins influence gene expression, particularly at imprinted loci. Using these tools we are able to understand more about the proteins that participate in the genomic landscape around imprinted loci and drive this unusual mode of gene regulation. These loci act as models for studying DNA binding proteins and their roles in transcription.

At the level of the individual locus, mechanisms of gene regulation were investigated using imprinted retrogenes as a model. Retrogenes are transcriptionally active intronless genes located within an intron of a 'host' gene. The role of epigenetic factors influencing gene transcription were investigated at the *H13/Mcts2* locus. Expression of an intronic retrogene can cause premature termination of a 'host' transcript. Our hypothesis is that this premature termination can be caused by transcription of the retrogene interfering with host gene transcription. We have designed a construct based on this locus, which allows us to regulate the expression through the retrogene to study this in more detail. These studies provide a mechanistic component to our whole genome analyses.

Acknowledgements

I would like to take this opportunity to thank everyone in the epigenetics research group for all of their help and support, both with the theoretical and practical aspects of this research. Special thanks go to Professor Rebecca Oakey for her support and guidance and for encouraging me to become a better scientist. I would also like to thank Dr Adam Prickett and Dr Mike Cowley for all of their help with the designing and execution of experiments, and the development of the project. I am grateful to Dr Reiner Schultz and Dr Nikolas Barkas for their help with the analysis of the ChIP-Seq data in this project.

I would like to thank Dr Bill Grey, Dr Nick Dand and Dr Rakhee Chauhan for all of their moral support over the past four years and for accompanying me on this journey, it would have been a much harder one without you. I would also like to thank Vanessa Haynes for her encouragement and support during the writing of this thesis.

Most of all I would like to thank Carl Johnson for his support and understanding over the past four years. I wouldn't have managed to finish without you.

I would also like to take this opportunity to thank the BBSRC Doctoral Training Grant for funding this research.

Contents

List of Figures

List of Tables

the mouse.

Abbreviations

3C – chromosome conformation capture

ADD - ATRX-Dnmt3-Dnmt3l domain

AID - activation-induced cytidine deaminase

APOBEC - apolipoprotein B mRNA-editing enzyme complex

ASE – alternatively spliced exon

ATP – adenosine triphosphate

ATRX - alpha-thalassemia/mental retardation, X-linked

bp – base pair

BxC – C57BL/6 x *Mus musculus castaneus*

BxJ - C57BL/6 x Japanese Fancy mouse F1

CdLS – Cornelia de Lange syndrome

ChIP-Seq – Chromatin immunoprecipitation sequencing

CPSF - cleavage and polyadenylation specificity factor

CstF - cleavage stimulatory factor

CTCF – CCCTC binding factor

CxB - *Mus musculus castaneus* x C57BL/6

DAXX - death domain-associated protein

DMR – Differential methylated region

Dnmt – DNA methyltransferases

dpc – days postcoitum

DSE – downstream sequence element

eGFP – enhanced green florescent protein

ES cells– Embryonic Stem cells

Frt - Flippase recognition target

Grb10 - growth-factor receptor bound protein 10

HATs – Histone Acetyltransferases

HDACs – Histone Deacetylases

HIRA – histone regulator A

HEK – human embryonic kidney

HKMT – Histone lysine methyltransferases

ICR – Imprinting control region

JxB – Japanese Fancy mouse F1 x C57BL/6

LCR – locus control region

LMP – low melting point

lncRNA – long non-coding RNA

MBD – Methyl binding domain

MCS – multiple cloning site

MeCP2 - Methyl CpG Binding Protein 2

miRNA - microRNA

nts - nucleotides

PAS – polyadenylation signal

PCR – polymerase chain reaction

Poly (A) - polyadenylation

PRMT – Protein arginine methyltransferases

qPCR - Quantitative Real-Time PCR

RNAi – RNA interference

SAM - S-adenosylmethionine

siRNA – small interefering RNA

SNP – Single Nucleotide Polymorphism

snRNA – small nuclear RNA

snoRNA – small nucleolar RNA

TADs – Topologically Associated Domains

TDG - thymine DNA glycosylase

Tet proteins - ten eleven translocation proteins

Tet-responsive – tetracycline responsive

TF's – Transcription factors

TRD – Transcriptional repression domain

TSS – Transcription Start Site

UCOE - Ubiquitously acting Chromatin Opening Element

USE – upstream sequence element

Chapter 1

Introduction

1.1. - Gene Expression

Gene expression is the process through which the information encoded in DNA is transcribed into RNA, before being translated into protein. Regions of DNA to be transcribed are marked by DNA binding proteins, which bind to specific regulatory sequences, and attract RNA polymerase, an enzyme that synthesises a strand of pre-RNA from the DNA template. As the RNA strand is being produced introns (and in some cases exons) are spliced out and the 5' (leading) end of the strand is capped. Once the stop codon of the gene has been transcribed the RNA strand is cleaved and a poly (A) tail is added in a process known as polyadenylation, to the 3' (trailing) end of the strand [1]. These modifications result in mature RNA, which is then exported from the nucleus to the cytoplasm, where it is translated into a protein. The expression of a particular gene can be regulated at any of these stages on the way from DNA to protein.

Not all RNA is translated into protein, roughly 98% of all transcripts in humans encode non-coding-RNA (ncRNA) [2]. There are many different types of non-coding RNA, thought to be involved in a wide range of processes in the cell. They include the classical RNAs like transfer RNA (tRNA), messenger RNA (mRNA) and ribosomal RNA (rRNA), as well as many newly discovered and less well-defined RNAs like microRNA (miRNA), small nuclear RNA (snRNA) and small nucleolar RNA (snoRNA). miRNAs are single stranded RNAs of about 22 nucleotides (nts) in length, generated from transcripts containing a local hairpin structure. They regulate transcription by binding to complimentary (but not completely complementary) regions

of mRNA transcripts, repressing their translation or regulating their degradation [3] [4]. miRNAs are involved in a wide range of developmental processes in both mammals and plants [2]. Short interfering RNAs (siRNAs) are similar to miRNAs, but they target messenger RNA to which they are completely complimentary, targeting them for degradation in the process known as RNA interference (RNAi) [5]. snRNAS are a small group of ncRNAs involved in splicing. snRNAs U1, U2, U4, U5 and U6 interact with over 200 proteins to form the spliceosome, and are responsible for splice-site and branch-site recognition, which are important for the removal of introns from the pre-mRNA [3] [6]. snoRNAs are between 60 and 300 nts in length, and tend to originate from the introns of protein coding or non-coding genes. They act to guide site-specific modification of other RNAs [5], and some snoRNAs have been shown to display tissue specific expression [2]. Long non-coding RNAs (lncRNA) are any RNAs over 100 nts in length, which are typically involved in imprinting and X-chromosome inactivation [7]. For example the lncRNA *Xist* (X inactive specific transcript) is essential for the silencing of the inactive X chromosome, through the recruitment of epigenetic factors associated with silencing to the inactive chromosome [8]. Piwi interacting RNAs (piRNAs) tend to be around 26 to 31 nucleotides in length, and are associated with the Piwi protein, which is involved in gametogenesis. They have been shown to regulate DNA methylation in germ cells in mice [9]. Enhancer RNA (eRNA) is thought to act as a scaffold to regulate the local chromatin architecture around their transcription start site, aiding expression through the gene [10].

1.2. – Imprinting

Diploid organisms have two copies of each autosomal chromosome, one inherited from each parent. Most genes are expressed from both parental alleles (biallelic expression).

However there are some genes that are only expressed from either the maternal or paternal allele (monoallelic expression, see Figure 1.1). These are known as imprinted genes and there are over 100 in the human genome [11].

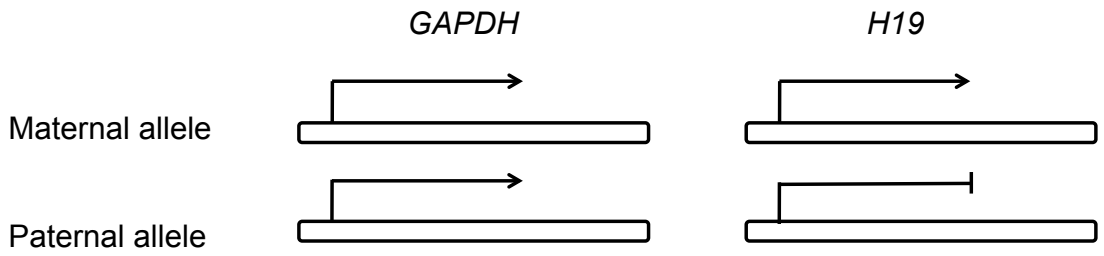


Figure 1.1 – Summary of biallelic vs monoallelic expression from a gene. *GAPDH* is an example of biallelically expressed gene, which is the case for most genes, it is expressed from both parental alleles. *H19* is an example of a monoallelically expressed gene, it is only expressed from the maternal allele and not the paternal allele.

The concept of imprinting was first described in 1984 by two independent studies published around the same time (McGrath, 1984 [12] and Surani, 1984 [13]). These groups both used pronuclear injection to generate murine embryos containing either two copies of the paternal genome (androgenetic) or the maternal genome (gynogenetic). Embryos containing both a maternal and a paternal copy of the genome, but generated through pronuclear injection were used as controls. The androgenetic and gynogenetic embryos failed to develop to full term. The androgenetic embryos developed to the 6-8 somite stage and showed extensive developmental impairment, while the trophoblast and extra-embryonic tissues were similar to controls but somewhat overgrown. The gynogenetic embryos appeared normal (they developed to the 25 somite stage) although smaller than controls, but with impaired development of the trophoblast. This is summarised in Figure 1.2. These experiments show that it is not the DNA sequence alone that is important for mammalian development. They illustrated that these genomes must be marked in different ways, so that gene expression from the paternal

genome is necessary for proper development of the extra-embryonic tissues and the maternal genome for the development of the embryo, and that both are required for normal embryonic development.

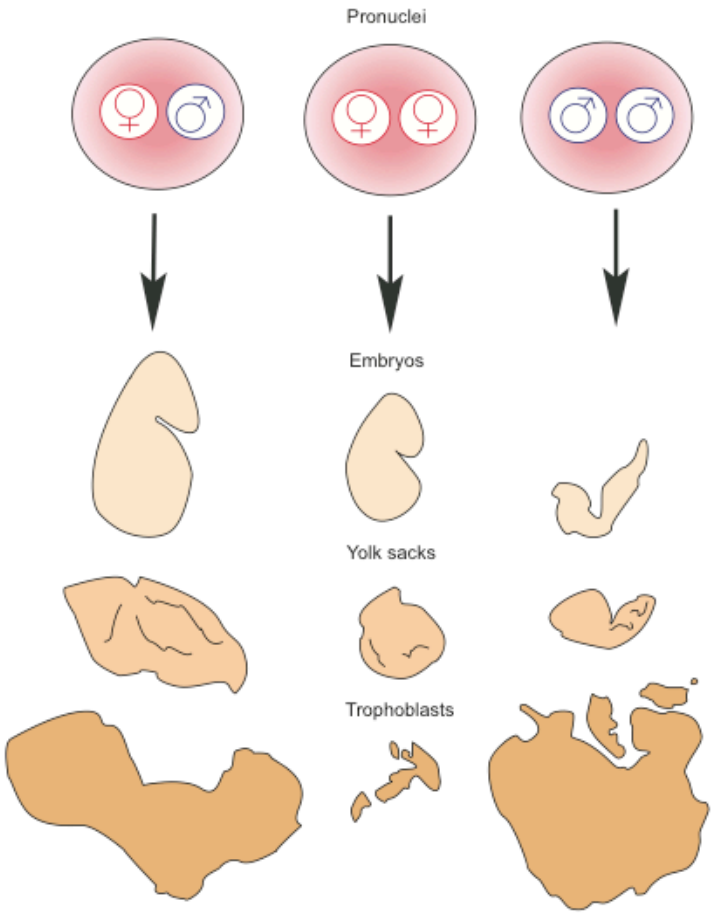


Figure 1.2 – The non-equivalence of the maternal and paternal genome, demonstrated through the generation of androgenetic and gynogenetic embryos (adapted from Surani, 1986 [14]).

Evidence for the non-equivalence of the parental genomes has also come from phenotypes observed during the breeding of reciprocal and Robertsonian translocation mice, at the MRC mouse breeding centre at Harwell [15] [16]. A Robertsonian translocation is a chromosomal abnormality in which two acrocentric chromosomes (where the chromosome arms are of different lengths) become joined by a common centromere [17]. For example intercrossing mice heterozygous for the Robertsonian

translocation Rb(11.13)4Bnr produced litters containing mice disomic for chromosome 11 or 13 (where both copies of the chromosome are inherited from one parent). While mice with disomy 13 develop normally, mice with disomy 11 show a size difference depending on which parent they inherit their two copies of chromosome 11 from. Mice with paternal chromosome 11 were much larger than their littermates, while those with maternal chromosome 11 were much smaller. Intercrosses with mice heterozygous for T(2;11)30H (the reciprocal translocation) showed that only the proximal region of chromosome 11 was involved. In this case, offspring with a maternal duplication or paternal deficiency of the proximal region of chromosome 11 were small at birth, having a similar phenotype to mice with maternal disomy 11. No offspring with a maternal deficiency or a paternal duplication of the proximal region of chromosome 11 were found, implying that these are lethal [18]. These translocation studies complemented the nuclear transfer experiments and lead to the identification of many imprinted genes.

The fact that both copies of an imprinted gene can be identical suggests the existence of another mechanism through which the two alleles can be differentiated, so that only one is expressed while the other is silenced. This mechanism must be heritable, as it must be maintained through mitosis throughout the life of the organism [19]. There is now a large body of work proposing that this mechanism is cytosine methylation, a type of epigenetic mark (reviewed by Ferguson-Smith, 2011 [17]).

A difference in the DNA methylation state between the two alleles of a gene can cause them to be expressed differently, making them imprinted. The region of DNA containing this difference in methylation is known as a differentially methylated region

(DMR), and this can be established either in the germline (a gDMR), or later in development in a somatic line. A DMR that has been experimentally shown to control the regulation of an imprinted gene in its vicinity is called an imprinting control region (ICR). The imprinted regions in mice are summarised in Table 1.1.

Gene/Cluster	Location	Methylated Allele	Gene	Allele expressed From
Gpr1	1qC2	P		P
			ZDBF2	P
Sfmbt2	2A1	M		P
Mcts2	2H1	M		P
			H13	M
Nnat	2H1	M		P
			Blcap	M
Gnas	2qH4	M		M
			Nesp	M
			Gnasxl	P
			Nespas	P
			Gnas exon 1A	P
Peg10	6qA1	M		P
			Sgce	P
			Ppp1r9a	M
			Asb4	M
Mest	6qA3	M		P
			Copg2	M
			Copg2as	P
			Klf14	M
			Nap1l5	P
Peg3	7qA1	M		P
			Zim2	M
			Zim1	M
			Usp29	P
			Zim3	M
			Zfp264	P
Snrpn	7qB5	M		P
			Atp10a	M
			Ube3a	M
			Snurf	P
			Ndn	P
			Magel2	P
			Mkrn3	P
			Peg12	P
			Ube3a-ATS	P
			snoRNAs	P
Cdkn1c	7qF5	M		M
			Kcnq1ot1	P
			Osbp15	M
			Nap1l4	M
			Phida2	M
			Slc22a18	M
			Msuit1	M

			Kcnq1	M
			Cd81	M
			Ascl2	M
			Tssc4	M
Igf2	7qF5	P		P
			Igf2as	P
			H19	M
			Ins2	P
Rasgrf1	9qE3.1	P		P
			Mir184	P
			AK029869	P
			A19	P
Plagl1	10qA2	M		P
			Phactr2	M
Grb10	11qA1	M		M
			Grb10as	M
			DDC	P
			Cobl	M
Zrsr1	11qA3	M		P
			Commd1	M
Dlk1	12qF1	P		P
			Rtl1	P
			Dio3	P
			Gtl2	M
			Anti-Rtl1 microRNAs	M
			Rian snoRNAs	M
			Mirg microRNAs	M
Peg13	15D3	M		P
			Kcnk9	M
			Trappc9	M
Igf2r	17qA1	M		M
			Pde10a	M
			Airn	P
			SLC22A2	M
			SLC22A3	M
Impact	18A2	M		M
Xlr3b	XqA7.3	M		M
			Xlr4b	M
			Xlr4c	M
Xist	XqD	M		P
			Tsix	M

Table 1.1 – Table of imprinted regions in mice (taken from the MRC Harwell Imprinting website [20], and the catalogue of imprinted genes [21] and Prickett, 2013 [22])

The *Ifg2-H19* locus is a classic example of an imprinted gene locus and is an excellent model from which to understand genomic imprinting and to test mechanisms of gene

regulation at other imprinted loci. At this locus the DMR is located at *H19* and is paternally methylated, meaning *H19* is only expressed from the un-methylated maternal allele. The promoter of *Igf2* is un-methylated on both alleles, nevertheless *Igf2* is imprinted; it is only expressed from the maternal allele [23]. This is due to the DMR at *H19*. On the un-methylated maternal allele, CTCF (an insulator protein that preferentially binds to un-methylated regions) binds to the DMR, blocking the interaction of an enhancer located downstream of *H19* with *Igf2*. On the paternal allele CTCF cannot bind due to the methylation at the DMR, allowing the interaction of the enhancer with *Igf2* and leading to *Igf2* expression (see Figure 1.3) [24].

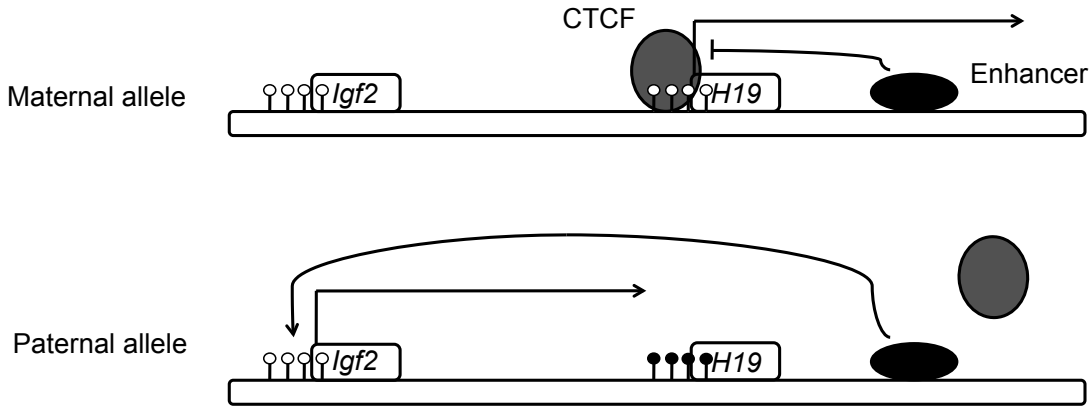


Figure 1.3 – Imprinted expression of *H19* and *Igf2*. The *H19* promoter is methylated (black circles) on the paternal allele, and un-methylated (white circles) on the maternal allele, where *H19* is expressed. The *Igf2* promoter is un-methylated on both alleles, but *Igf2* is only expressed from the paternal allele, as CTCF (grey oval) bound at the DMR of *H19* blocks the interaction between *Igf2* and its enhancer (black oval). The boxes represent the promoter of the gene, and not the entire gene.

There is evidence to suggest that CTCF binding to the maternally un-methylated DMR during development acts to protect the DMR from gaining methylation. RNA interference (RNAi) was used to knock down the levels of CTCF in the mouse oocyte. Of the five founder females generated the three showing the greatest reduction in CTCF mRNA and protein were hypermethylated at the *H19* DMR, compared to the non-transgenic control females, which were hypomethylated at the *H19* DMR [25]. These

experiments show that CTCF is required during development to maintain the hypomethylated status of the maternal *H19* DMR.

1.3. - Mechanisms of Gene Expression

Some genes only need to be expressed for short periods of time; for a defined stage of development or in response to changes in their environment for example. As expression from such a gene is essential during these periods but can be detrimental at other times, the cell has devised a number of ways to regulate gene expression, and expression from a gene can be controlled through several of these mechanisms.

1.3.1. - Regulatory Elements

These are sequences of DNA that can alter transcription through a gene. Promoters are usually located at the transcription start site (TSS) of the gene and directly activate transcription by allowing the binding of transcription factors and RNA polymerase II. Many genes also require cis-regulatory elements, which can be located up to 1Mb away from the TSS (either upstream or downstream), and can be located within introns of neighbouring genes [26]. Enhancers are a type of cis-regulatory element that act to stimulate transcription by recruiting tissue-specific transcription factors (TFs), RNA polymerase II and other cofactors involved in transcriptional activation [27]. Repressors are a type of cis-regulatory element that recruit proteins that decrease the expression of a gene or inactivate it [28]. The final type of cis-regulatory element are insulators, which recruit proteins that isolate regions of DNA or regulatory elements from each other [28]. For example, if located between the TSS of a gene and an enhancer, an insulator can block the activity of the enhancer so that it has no effect on expression from the gene.

Topologically associating domains (TADs) can act as insulators to restrict interactions betweens genes, promoters and enhancers [29]. TADs are regions of about 100kb, where the genes they contain form more interactions with each other, than they do with genes in neighbouring TADs [30]. The boundaries between TADs are usually associated with binding sites for the insulator protein CTCF, and appear to stop the spread of heterochromatin between TADs [30]. TADs are conserved across species and cell types providing support of the important role they play in the regulation of transcription. There is evidence that disruption of the boundaries between TADs can cause misregulation of genes and can lead to disease [29]. Liebenberg syndrome is an autosomal dominant disorder where an individuals arms exhibit morphological characteristics of their legs. This is due to the deletion of a boundary element, which usually protects *PITX1* (a gene involved in hindlimb development) from the activity of a neighbouring enhancer element. When the boundary is removed the enhancer can activate expression from *PITX1* [29]. Disruptions to the TAD boundaries flanking *EPHA4* (a receptor tyrosine kinase) can lead to brachydactyly (short digits) or syndactyly (the fusing together of digits) depending on which boundary is disrupted, despite *EPHA4* having no involvement with limb development. A disruption to the TAD boundary on the telomeric side of *EPHA4* allows limb enhancers to interact with *PAX3* (which is usually located in a neighbouring TAD), causing it to be incorrectly expressed and resulting in brachydactyly. A disruption to the TAD boundary on the centromeric side of *EPHA4* allows the limb enhancers to interact with *WNT6* (usually located in a neighbouring TAD), again causing it to be incorrectly expressed, resulting in syndactyly [31] [29]. There is also evidence that disruption of TAD boundaries can play a role in the progression of cancer [29].

1.3.2. – DNA Methylation

DNA methylation is the addition of a methyl group to the 5' position of a cytosine residue (Figure 1.4). This mainly occurs when the cytosine is located next to a guanine in the DNA sequence of mammals, known as a CpG dinucleotide, but can occur on a cytosine not located next to a guanine. CpG dinucleotides occur throughout the genome, but are also clustered in regions known as CpG islands [32]. CpG islands are defined as regions of at least 200bp with a CG content of over 50% and with an observed-to-expected CpG ratio of over 0.6 (for most of the genome the number of CpG dinucleotides is much lower than 0.6 times what would be statistically expected) [33]. Isolated CpG dinucleotides tend to be methylated, whereas those located in CpG islands tend to be un-methylated [32]. CpG islands are associated with about 70% of mammalian genes including most housekeeping genes and a large proportion of tissue specific genes [34], usually at or near the promoter. Most cytosines in a CpG island are un-methylated unless the gene is silenced, in which case they can be methylated on both alleles. However, allele-specific methylation is a special case of gene regulation where one parental allele is methylated and the other un-methylated. This is the case for imprinted genes.

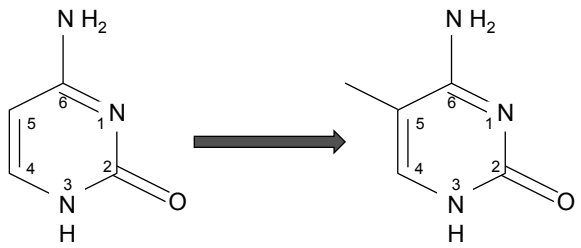


Figure 1.4 - Methylation occurs at the 5' carbon of cytosine.

Methylation of a cytosine located next to an adenosine, cytosine or thymine is known as non-CpG methylation. Non-CpG methylation occurs most commonly at CpA dinucleotides, less commonly at CpT dinucleotides and infrequently at CpC dinucleotides [35]. The distribution and abundance of non-CpG methylation varies across the genome and different cell types, but is more abundant in pluripotent cells than in most differentiated cells (high levels of non-CpG methylation have found in brain tissue) [36]. For example about 25% of methylated cytosines in human ES cells are located in non-CpG dinucleotides, compared to less than 1% in human fibroblasts [37]. There is evidence that non-CpG methylation at promoters is associated with reduced gene expression. Non-CpG methylation at the promoter of the *B29* gene in human B cells inhibits the binding of the early B cell factor transcription factor, causing a reduction in the activity of the promoter [38]. Similarly the presence of non-CpG methylation in the promoter of the *syt11* gene (human synaptotagmin XI, which has been associated with the development of schizophrenia) blocks the binding of Sp family proteins, reducing the activity of the promoter, and expression through the gene [39]. The expression of *PGC-1α* (which is involved in mitochondrial function) has also been shown to be influenced by non-CpG methylation in the skeletal muscle of patients with Type II Diabetes Mellitus. Again non-CpG methylation at the promoter of *PGC-1α* leads to a decrease in expression of the gene. However, this methylation can be altered in response to other factors in the cell, as *TNF-α* and free fatty acids can induce non-CpG methylation [40]. Non-CpG methylation may also play a role in the regulation of imprinted genes, as it is present on the repressed allele of several imprinted genes, including *Peg13*, *Sgce*, *Grb10* and *Kcnq1ot1*, in adult mouse frontal cortex [41]. However, more research is needed to elucidate the role non-CpG methylation plays in the regulation of gene expression.

1.3.2.1. – Mechanisms of DNA Methylation

Although DNA methylation is a stable mark, in the sense that it is maintained during fertilisation and cell division, it is also dynamic; the methylation status can be altered when required (for example during germ cell development). The addition of a methyl group is catalysed by a family of DNA methyltransferases (Dnmts) (see Table 1.2 for more detail). Dnmt3a and Dnmt3b are the *de novo* Dnmts as they can establish new patterns of methylation on DNA strands. Dnmt1 is the maintenance Dnmt as it maintains current patterns of methylation during DNA replication by copying the methylation pattern from the parental strand to the new daughter strand. It also has the ability to repair DNA methylation [42]. The activity of Dnmt3a and Dnmt3b can be modulated by association with another family member; Dnmt3l, which lacks a catalytic domain and therefore acts as a cofactor.

Dnmt3l is expressed during development and is required for the establishment of genomic imprinting [43]. Bourc'his *et al* demonstrated that disrupting both copies of the *Dnmt3l* allele in mice resulted in sterility. It was found that heterozygous progeny of homozygous mothers failed to develop past 9.5 days postcoitum (dpc), due to abnormalities in the extraembryonic tissues. In addition, bisulphite genomic sequencing of the embryo DNA showed that the maternally imprinted genes *Snrpn* and *Peg1,* were un-methylated on both alleles. *H19*, a paternally imprinted gene, showed the expected allele-specific methylation [43]. These results show that Dnmt3l is required for the establishment of maternal imprinting during oogenesis.

	Function	Reference
Dnmt1	Maintenance of methylation during replication	Moore, 2013 [42]
	Essential for the survival of foetal mitotic neuroblasts Role in organ development Cellular differentiation	Fan *et al*, 2001
	Genomic stability Establishing intestinal epithelial crypts after birth	Elliott *et al*, 2015 [44]
Dnmt3a	Establishment of new sites of methylation	Bourc'his, 2001 [43]
	Methylation of CpT, CpA and CpC sites (cytosine-phosphate-thymine / adenosine / cytosine respectively)	Uysal, 2015 [45]
Dnmt3b	Establishment of new sites of methylation	Bourc'his, 2001 [43]
	Methylation at repeated DNA sequences in the pericentric satellite regions of chromosomes	Okano, 1999 [46]
Dnmt3l	Establishment of genomic imprinting	Bourc'his, 2001 [43]
	Essential for fertility	Bourc'his, 2004 [47]

Table 1.2 – Dnmt family members and their role in methylation.

1.3.2.2. – Mechanisms of DNA Demethylation

In order for DNA methylation to be a dynamic mark it must also be possible to remove the methyl group from the cytosine. Demethylation can occur either passively, through the inhibition or repression of Dnmt1 during replication, or actively. No direct DNA demethylases have been identified, which suggests that methyl groups are removed indirectly. This is thought to occur through the base excision repair pathway, which can remove a methylated cytosine and replace it with an unmodified cytosine. There are two ways the base excision repair pathway can be triggered. Firstly, methylated cytosine can be deaminated by activation-induced cytidine deaminase (AID) or apolipoprotein B mRNA-editing enzyme complex (APOBEC) to form thymine, which is recognised as being a mismatched base and removed by thymine DNA glycosylase

(TDG), a member of the base excision repair pathway [42]. Alternatively, a family of hydroxylases have recently been implicated in this indirect demethylation pathway. The Tet (ten eleven translocation) proteins, in the presence of their cofactors Fe^{2+} and 2-oxoglutarate, can oxidise methylated cytosine through a series of intermediates to form 5-carboxy-cytosine (see Figure 1.5) [48] [49]. 5-carboxy-cytosine can then be removed by TDG [42], as above.

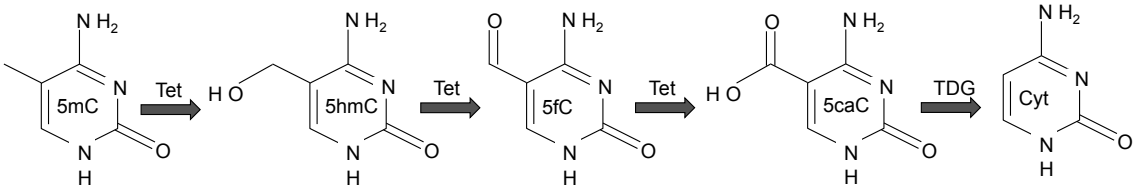


Figure 1.5 – Mechanism of active demethylation. Methylated cytosine is oxidised by Tet enzymes to generate 5-hydroxymethyl-cytosine (5hmC), which is further oxidised to form 5-formyl-cytosine (5fC). This in turn is oxidised to form 5-carboxy-cytosine (5caC), which is recognised by TDG (thymine DNA glycosylase), a component of the base excision repair pathway, and the base is cleaved from the phosphate backbone and replaced with cytosine.

In mammals genomewide demethylation takes place early in the development of the germline. By the time mature sperm and oocytes have developed, the level of methylation in the genome will have risen to the levels seen in somatic cells. This wave of rapid demethylation of the genome (complete by embryonic day 14) followed by a slower re-methylation, allows the removal of any acquired epigenetic modifications, and the correct resetting of methylation patterns. The expression of Dnmts corresponds to the changing methylation patterns in these germ cells [50] (see Figure 1.6). There is little or no expression of any of the Dnmts while the genome undergoes demethylation. However, as the expression of the Dnmts increases, methylation of the genome also starts to increase.

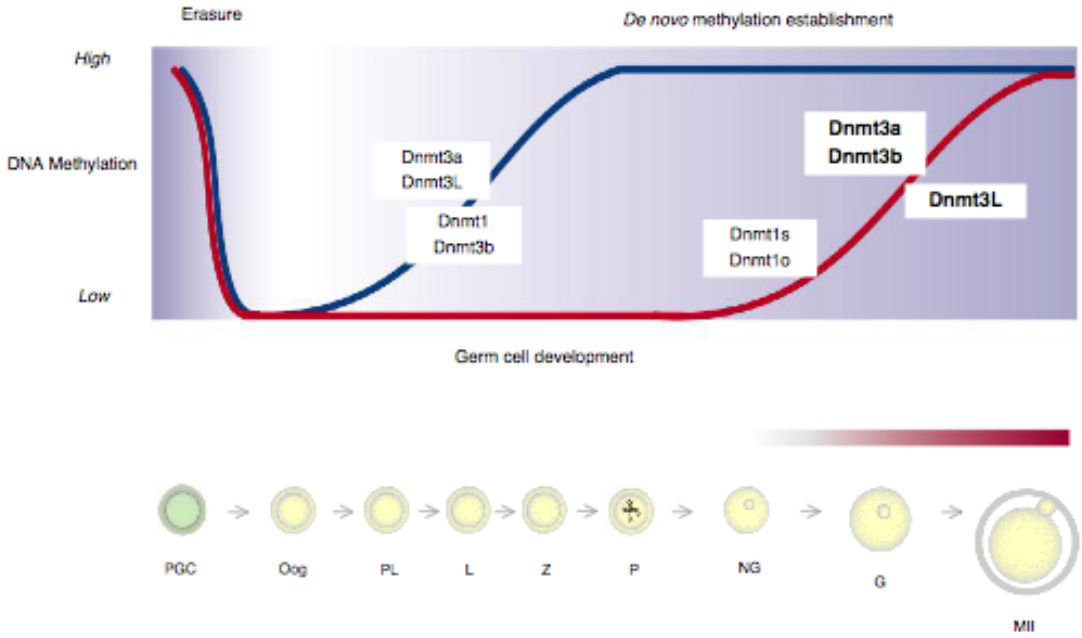


Figure 1.6 – The changing levels of DNA methylation and expression of DNA methyltransferases during mammalian germ cell development. The blue line represents methylation of paternally imprinted genes and the red line represents methylation of maternally imprinted genes. Methylation of non-imprinted genes closely follows the methylation of imprinted genes. The expression of the Dnmts is shown. Developmental stages - PGC = primordial germ cell, Oog = oogonia, PL = preleptotene, L = leptotene, Z = zygotene, P = pachytene, NG = non-growing oocyte, MII = metaphase II oocyte (reproduced from Lucifero, 2007 [51]).

1.3.2.3. – Non-methyl Cytosine Modifications

There is evidence that the intermediates generated during the demethylation of methylated cytosines could be viewed as cytosine modifications in their own right. 5hmC is the most common of these modifications, occurring at varying levels in different tissue types, and being most abundant in brain. In the cortex of the human brain 5hmC occurs at about 1% of all cytosine residues, or at 20-25% of all 5mC bases [52]. 5hmC tends to be located at promoters or within gene bodies, and has been associated with gene activity [52], but this localisation varies across cell types. In ES cells 5hmC tends to be localised to enhancers, but in neurons it is located within the gene bodies of genes required for the functioning of the neuron. 5hmC seems to be

associated with active genes [53]. Levels of 5hmC have been shown to be high in pluripotent cell types [53], and reduced in cancer cells compared to their normal neighbours [52]. 5fC and 5caC are present at much lower levels, about 10-1000 fold lower than the levels of 5hmC, across all cell types. 5fC is present at distal regulatory elements, preferentially at poised enhancers or the promoters of poised genes [53]. It is thought that these three modifications could play a role in gene regulation.

These modifications can affect the activity of certain restriction enzymes, and so restriction digests can be used to determine the presence of modified cytosines in a sequence. *Msp*I can digest 5C, 5mC and 5hmC, but not 5fC or 5caC [49]. *Msp*I and *Hpa*II (its isoschizomer) can be used to determine the presence of 5hmC in a defined sequence. These two restriction enzymes have differing sensitivities to the glycosylation of 5hmC, so treating the sample with β-glucosyltransferase (βGT) inhibits the activity of *Msp*I, but not *Hpa*II. To determine the abundance of 5hmC the sample is treated with βGT and then digested with either *Msp*I or *Hpa*II, amplified using PCR, and then the results are compared [54]. Alternatively after the sample has been treated with βGT it can be digested with *Aba*SI (which recognises glycosylated 5hmC, but not other cytosine modifications), which digests the DNA about 11-13nts or 9-11nts away from the 3' end of the glycosylated 5hmC. The DNA fragments are then sequenced and mapped back to the genome to determine the location of 5hmC [55]. *MspJ*I and *PvuRts*I can also be used to discriminate between 5mC and 5hmC, as they will only digest 5hmC or glycosylated 5hmC, and not 5mC [56].

There are several methods available for the genomewide detection of these modifications, and they all utilise bisulphite conversion. Sodium bisulphite can be used

to deaminate cytosine generating a uracil base. The presence of a methyl group on a cytosine protects it from this deamination. When the sodium bisulphite treated DNA is then amplified in a PCR reaction, the uracil (from the deaminated cytosine) is amplified as a thymine, while the methylated cytosines are amplified as cytosine. Any cytosines in the resulting sequence show the presence of methylated cytosines in the original sequence [57]. The bisulphite treated sample can then be used for pyrosequencing (where the region of interest is amplified through PCR [58]) or whole genome sequencing (BS-Seq). This protocol has formed the basis of the sequencing methods used to detect 5hmC, 5fC and 5acC. Tet-assisted Bisulphite Sequencing (TAB-Seq) and Oxidative Bisulphite Sequencing (oxBS-Seq) can differentiate between 5mC and 5hmC [59], which are both detected by BS-Seq. In TAB-Seq βGT is used to add a glucose group onto 5hmC, to generate β-glucosyl-5-hydroxymethylcytosine (5gmC), which protects the 5hmC from oxidation by TET. The sample is then treated with TET, which oxidises all the 5mC to 5caC, before being treated with sodium bisulphite, which converts all the cytosines and 5caC to uracil or 5caU but leaves the 5gmC. When the sample is sequenced the cytosines in the sequence will show the location of 5hmC in the original sequence [60]. In oxBS-Seq potassium perruthenate is used to oxidise 5hmC to form 5fC, which upon treatment with sodium bisulphite is deaminated to form uracil. In this experiment only 5mC is protected from deamination by sodium bisulphite and so when the sample is sequenced the cytosines show the location of 5mC. To determine the location of 5hmC BS-Seq (which detects both 5mC and 5hmC) must be performed and the results compared to those generated from the oxBS-Seq [61]. 5caC can be detected through Chemical Modification-assisted Bisulphite Sequencing (CAB-Seq), which uses 1-ethyl-3-(3-dimethylaminopropyl)carbodiimide hydrochloride (EDC) to attach an amine group to 5caC, protecting it against deamination by sodium

bisulphite. In the resulting sequence cytosines mark the location of either 5mC, 5hmC or 5caC. To determine the location of 5caC only these results must be compared BS-Seq where 5mC and 5hmC are read as cytosines and 5caC is read as a thymine [62]. Reduced Bisulphite Sequencing (redBS-Seq) can be used to detect 5fC. Sodium borohydride is used to reduce 5fC to 5hmC, before bisulphite treatment and sequencing. 5hmC protects the cytosines from deamination by sodium bisulphite and so any cytosines in the resulting sequence show the presence of 5mC, 5hmC or 5fC. To determine the location of 5fC in the original sequence the redBS-Seq results are compared to BS-Seq, where only 5mC and 5hmC are read as cytosines, and 5fC is read as a thymine [63].

1.3.3. - DNA Binding Proteins

There are many different families of proteins that can bind to DNA and affect expression of a gene.

The transcription factors are a family of DNA binding proteins involved in gene transcription. These bind to specific sequence motifs in the promoter or enhancer of the gene and act to recruit RNA polymerase to the TSS [64] [65].

There are only three forms of RNA polymerase, making it a very small family of DNA binding proteins, but as they transcribe DNA into RNA they are an important one [66].

Histones are another important group. These are proteins around which DNA is coiled to form chromatin, allowing the whole genome to be packed into the small space of the nucleus. Gene expression can be regulated through this coiling, with tightly coiled

chromatin (heterochromatin) repressing expression, and more open chromatin (euchromatin) allowing expression of genes within these regions [67].

1.3.3.1. – Methylation Sensitive Binding Proteins

The methyl-CpG-Binding Domain (MBD) family bind preferentially to methylated DNA, except for MBD3 which has a similar binding affinity for both methylated and unmethylated CpG dinucleotides and MBD5 and MBD6 which bind unmethylated cytosine [68] [69] [70]. The MBD folds to create a wedge shaped structure, which recognises methylated CpG dinucleotides [68]. Apart from this domain there is little structural similarity between family members, although MBD2 and MBD3 do have very similar DNA sequences [68], and several family members, including MeCP2 which I shall discuss further in section 1.3.3.6., contain a transcriptional repressor domain (TRD) [70]. These proteins play an important role in methylation dependant gene regulation [69], and have been implicated in neurogenetic disorders and various cancers [70].

MBD proteins play an important role in transcriptional regulation as they link methylation marks to histone modifications, which can alter the chromatin landscape around certain genes to influence their expression [70]. Therefore any disruption to their normal functioning can have wide-ranging implications. Most of the MBD proteins have been associated with neurological disorders. Mutations in MeCP2 play a role in the development of Rett syndrome [71], which will be described in more detail in section 1.3.3.6. Missense mutations in MBD4, and mutations in the MBD of MBD4 have been implicated in the development of autism spectrum disorders [70] [72]. SNP mutations and single base insertions in MBD5 and MBD6 have also been implicated in

the development of autism spectrum disorders [70], possibly through the disregulation of the development and proliferation of neural cells due to the altered function of these proteins [72].

Mutations in MBD proteins, as well as their overexpression or inhibition have been associated with multiple cancer types [70] [72]. For example overexpression of MBD2 has been associated with the hypermethylation of the promoter of *GSTP1*, a well-studied tumour suppressor gene, in prostate cancer [70]. Knockdown of MBD2 has been shown to remove the repression of several tumour suppressor genes like $p16^{INK4a}$ and $p14^{ARF}$, increasing the possibility of developing cancer [70]. Frameshift mutations in MBD4 have been associated with the development of colorectal, gastrointestinal, pancreatic and endometrial cancers [70].

1.3.3.2. – Non-methylation Sensitive Binding Proteins

Just as there is a family of proteins that recognise and bind to methylated CpG dinucleotides, there is also a family of proteins that preferentially bind to unmethylated CpG dinucleotides. These proteins contain a 35-42 amino acid zinc finger CxxC domain (ZF-CxxC) [73], which recognises unmethylated CpG dinucleotides and therefore CpG islands. Proteins containing this domain are involved in the regulation of chromatin modifications and DNA methylation, which suggests that they could play a role in defining the chromatin environment of CpG islands [74]. For example CxxC finger protein 1 (CFP1) forms part of a methyltransferase complex with SETD1, which is responsible for the methylation of lysine 4 of histone H3 (H3K4me3) [75]. This mark is usually associated with the activation or poising of genes, and has been shown to block the activity of Dnmt3L, protecting the region it marks from *de novo* methylation

[76]. Regions containing large quantities of unmethylated CpG dinucleotides, like CpG islands, are sufficient to recruit CFP1 [77] leading to the deposition of the H3K4me3 mark, to block *de novo* methylation and maintain a chromatin state allowing the initiation of transcription [75].

KDM2A is a lysine demethylase containing a ZF-CxxC domain, which acts to remove the methyl groups from lysine 36 on histone H3 over the CpG islands where KDM2A binds [74]. The function of the H3K26me2 is poorly defined, but studies in yeast imply that it could play a role in the recruitment of EAF3, which is part of the RPD3S histone deacetylase complex, to repress transcription [75]. Therefore removing this mark from CpG islands could act to create a chromatin environment that permits initiation of transcription.

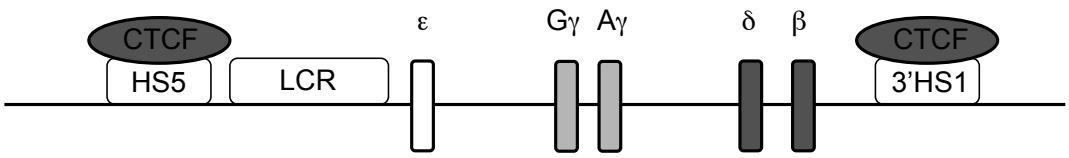
In sections 1.3.3.3 – 6. I will discuss further the four DNA binding proteins that are relevant to this thesis.

1.3.3.3. – CTCF

CTCF is an 11-zinc finger CCCTC DNA binding protein that is highly conserved across all vertebrates. The function of the zinc fingers is to bind CTCF to DNA [78]. It has a consensus binding sequence (as the binding site varies slightly in length and sequence between tissues) of about 11-15bp; this sequence is bound by 4-5 central zinc fingers. Variations at the 12 bp of the consensus binding sequence can alter the methylation of the binding site, altering CTCF binding [79]. CTCF forms transcriptional boundaries and has been extensively studied in the context of imprinted genes [11], as it binds preferentially to un-methylated regions of the genome. It plays an important role at the

imprinted *Igf2-H19* locus, binding to the un-methylated DMR on the maternal allele to regulate the expression of *Igf2* by blocking its interaction with the enhancer (for more detail refer back to section 1.2 and Figure 1.3). CTCF also plays a key role at the β-globin locus. In humans and mice this locus is flanked by CTCF sites, which were thought to insulate it from the surrounding chromatin [80]. However, chromosome conformation capture (3C) experiments have shown that these CTCF sites interact with each other, forming large loops of DNA in erythroid progenitor cells (Figure 1.7). One of these loops contains the locus control region (LCR), main regulatory elements and the genes [80], bringing these elements into close enough proximity to allow transcription of these genes to occur.

A



B

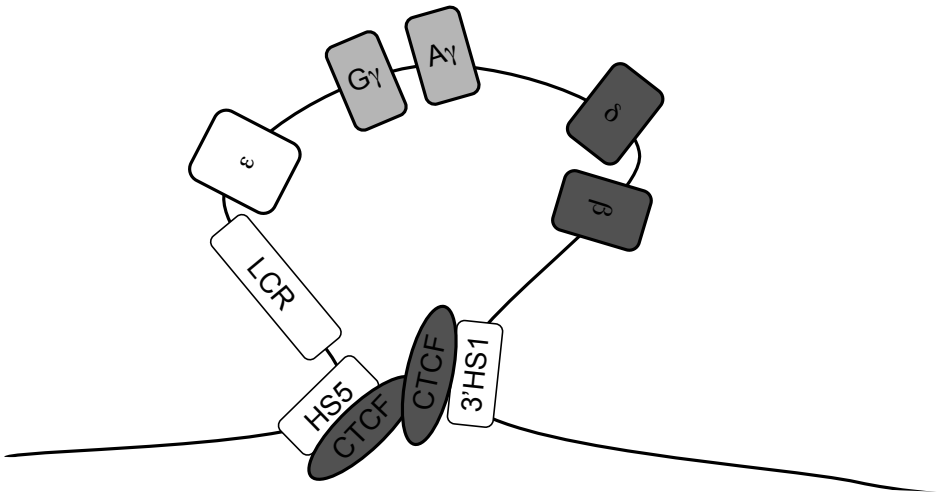


Figure 1.7 – Overview of the β-globin locus and its interactions with CTCF. A- Diagram of the β-globin locus in humans. B – Chromatin loop formation at the β-globin locus. HS5 and 3'HS1 are insulator elements, which CTCF binds to. ε is expressed in the embryo, Gγ and Aγ are expressed in the foetus, and δ β are expressed in adults. Adapted from (Kim, 2012 [81]).

CTCF plays a role in the formation of chromatin loops in other regions of the genome, forming and stabilising interactions between distant regions of the chromosome [82], and regulating the expression of genes located in these regions. CTCF has been shown to co-localise with cohesin during the formation of these loops [78].

1.3.3.4. – Cohesin

The cohesin complex is made of four core subunits; SMC1α, SMC3, Rad21 and SA1/SA2, which form a ring (Figure 1.8). It is involved in many chromosomal processes including double-strand break repair, condensation (the reorganisation of extended chromatin chains into compact chromosomes during meiosis and mitosis) and

gene expression, as well as its traditional role in maintaining the attachment of sister chromatids during mitosis [83]. The cohesin ring opens to embrace the two chromatids and hold them together [84]. This ability to hold sister chromatids together is also essential for homologous recombination for the repair of double-stranded DNA breaks in G2 (the second growth phase of the cell cycle, where the cell prepares for cell division). By bringing the sister chromatids together the intact chromatid can be used as a reference for the repair of the broken one [85].

Cohesin binding sites have been found at the promoters of active mammalian genes [86], implying a role for cohesin in gene expression. 3C experiments have shown that cohesin is also involved in stabilising or forming long-range interactions between its binding sites [86]. Cohesin can also act to regulate gene expression during interphase, through binding with CTCF [87] [88].

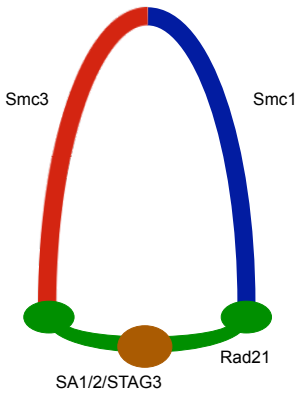


Figure 1.8 – Diagram of the structure of cohesin. Adapted from (Mehta, 2013 [89]) Smc1 and Smc3 are long antiparallel coiled-coil structures that heterodimerise at one end. The other ends are joined by Rad21 to close the ring. SA1/SA2 interacts with the ring through Rad21 [84].

As cohesin is involved in several essential processes in the cell, mutations in this complex lead to at least three diseases in humans, known collectively as the cohesinopathies; Cornelia de Lange syndrome, Roberts syndrome and the Warsaw

Breakage syndrome [90]. Cornelia de Lange syndrome (CdLS) is an autosomal dominant disease affecting 1 in 10,000 individuals [83], while the other cohesinopathies are recessive and much rarer. Mutations in cohesin and its interacting proteins are also implicated in several cancers [90].

CdLS and the other cohesinopathies are characterised by craniofacial and limb deformities, developmental abnormalities and mental retardation [91] [83]. 65% of CdLS cases are caused by mutations in *NIPBL,* whose product forms a heterodimer with MAU2 which is involved in loading cohesin onto chromatin. Mutations in the SMC subunits of cohesin account for around 7% of CdLS cases (5% are due to *SMC1α* and 1-2% to *SMC3*). Around 5% of cases are caused by mutations in *HDAC8*, causing a loss of HDAC8 activity [83]. HDAC8 is a SMC3 deacetylase, which is essential for the deacetylation of SMC3 after sister chromatid cohesion so that it can be re-loaded onto chromatin for the next cycle. Acetylated cohesin is loaded onto chromatin at a reduced rate negatively affecting transcription [89].

1.3.3.5. – ATRX

The ATRX (alpha-thalassemia/mental retardation, X-linked) gene is located on the X chromosome. It consists of 36 exons, spanning about 300kb, in humans. At least two alternatively spliced transcripts are generated from the gene differing at their 5' ends, and producing slightly different proteins. A shorter transcript, which retains intron 11 and then terminates, produces a truncated version of the protein, known as ATRXt [92]. ATRX is involved in many different cellular processes, and as such generating a knock-out mouse has proved difficult. A traditional knock-out of ATRX is lethal, and so

conditional knock-outs, where ATRX is knocked-out in a particular tissue, must be used [93] [94].

ATRX is a member of the SW12/SNF family of proteins, which are involved in chromatin remodelling. ATRX has been shown to be involved with gene silencing. It interacts with death domain-associated protein (DAXX) to deposit H3.3. H3.3 is a histone, which is essential for mammalian development; it is usually associated with transcription in which case it is recruited to gene rich regions by histone regulator A (HIRA). However ATRX and DAXX act to deposit H3.3 at telomeres, pericentric heterochromatin and at the methylated allele of a DMR [95]. In the latter case this causes gene silencing by blocking the recruitment of the necessary transcriptional machinery (Figure 1.9).

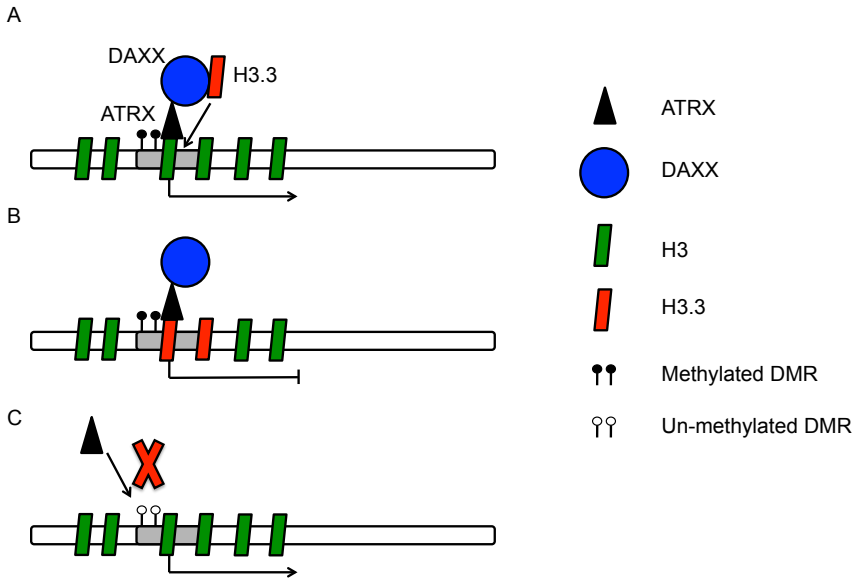


Figure 1.9 – Schematic of ATRX mediated recruitment of H3.3 to DMRs. A and B – ATRX recruits DAXX to the methylated allele of a DMR. DAXX deposits the histone variant H3.3 onto chromatin at the DMR, to repress transcription through the allele. C – ATRX is unable to bind to the un-methylated allele of a DMR, therefore DAXX and H3.3 cannot be recruited allowing transcription to occur through the allele. Modified from (Voon 2015) [95].

ATRX contains a plant homeodomain zinc finger (PHD, a chromatin interaction motif), which due to its similarity with a region found in the Dnmts has been designated the ATRX-Dnmt3-Dnmt3l domain (ADD), and an ATPase/helicase domain [96]. The SW12/SNF family of proteins use their ATPase/helicase domains to slide histones along the chromatin, remodel them or remove them completely [97].

Mutations in the ATRX gene are associated with alpha-thalassemia/mental retardation, X-linked syndrome, and a range of phenotypically similar syndromes including Carpenter-Waziri syndrome, Holmes-Gang Syndrome, Juberg-Marsidi Syndrome and Smith-Fineman-Myers Syndrome [92], which are predominately found in males. As well as alpha-thalassemia, symptoms also include mental retardation, facial, skeletal and urogenital abnormalities [97]. 113 different disease associated mutations have been identified in the ATRX gene so far. The mutations tend to be missense mutations, and cluster in the ADD domain (about 50%) and the ATPase/helicase domain (about 30%) [92].

1.3.3.6. – MeCP2

MeCP2 (Methyl CpG Binding Protein 2) binds to methylated CpG dinucleotides located within 11bp of a run of at least four adenine/thymine nucleotides [98], and can act to recruit the helicase domain of ATRX to heterochromatic foci in a DNA dependent manner [71]. MeCP2 is located on the X chromosome and mutations in this gene are associated with Rett Syndrome, a neurodevelopmental disorder [71]. Rett syndrome occurs in about 1 in 10, 000 female births. Development proceeds normally until 6 to 12 months of age when symptoms start to appear and there is a loss of learned skills, like speech and controlled movement. Other symptoms include social withdrawal,

respiratory problems, deceleration of head growth, intellectual impairment, and seizures [99] [100] [101].

MeCP2 contains three functional domains; MBD at the N-terminus, a nuclear localisation sequence and a TRD. There are currently 1,013 documented mutations spread through these three domains [102]. The eight most frequent mutations are C to T conversions. All types of mutation can be found throughout the gene, with frameshift mutations tending to cluster in the C-terminal domain and the TRD, and missense mutations in the MBD [103].

1.3.4. - Histones

The basic unit of chromatin is the nucleosome. DNA is wrapped around an octamer of four core histones, with two copies of each histone present. These core histones are H2A, H2B, H3 and H4. About 147bp of DNA is wrapped around each octamer and together these comprise a nucleosome [67]. The region of DNA between each nucleosome is termed the linker region and can range between 10bp and 100bp in length. Linker histones (H1) bind to these regions and play a role in chromatin folding [104].

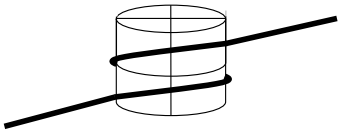


Figure 1.10 – Diagram of a nucleosome. About 147bp of DNA is wrapped around an octamer comprising of two copies of each of the following histones; H2A, H2B, H3 and H4.

Histones are essential for the correct folding of the DNA into chromosomes. The histone genes are located in three main clusters in humans: HIST1 consists of 55 genes

on chromosome 6, HIST2 contains six genes and is located on chromosome 1 and HIST3 containing three genes on chromosome 1. There is a lot of redundancy amongst the histone genes, for example there are 14 genes that encode the same H4 protein [105]. As well as there being multiple copies of each gene there are also variants of all of the histone proteins except for histone 4. There are 11 variants of histone 1, with six of these being tissue-specific. For example H1t and H1T2 are only expressed in the testis [106].

There are at least 19 variants of histone H2A, encoded by 26 genes [107]. Only about four of these have been studied in detail; H2A.X, H2A.Z, macroHA and H2ABBD [108]. H2A.X is the most commonly occurring variant [109], and is essential for DNA repair and recombination [108]. H2A.Z is another important variant being essential for embryogenesis as it plays a role in establishing pluripotency [110], as well as playing roles in transcription and suppression of antisense RNA [109]. macroH2A is involved in X chromosome inactivation and repression of transcription [108] [109], while H2ABBD (H2A Barr-body-deficient) plays a role in spermatogenesis [109]. 16 distinct variants of histone H2B have been identified [111], but very little is known about their functions. Three of the variants (H2B.1, H2B.W and subH2B) are all involved in spermatogenesis [109]. There are six histone H3 variants; centromeric H3 (or cenH3), H3.1, H3.1t (also known as H3.4), H3.2, H3.3 and H3.3C, encoded by a total of 18 genes [107]. These variants have specific roles, for example H3.3 is enriched at transcriptionally active genes and regulatory elements [112], cenH3 is essential for kinetochore assembly [108] and H3.1t is expressed only in the testis [113].

1.3.4.1. – Histone Modifications

The N-terminal tail of each histone extends out from the nucleosome, and can be modified in a variety of ways. Each histone can be subject to different modifications at the same time. The different combinations of modifications are known as the 'histone code', which can be 'read' by other proteins to affect transcription [114].

One of the most common modifications is methylation, which mainly occurs on lysine and arginine residues of the tail. Lysines can be mono-, di- or tri-methylated, whereas arginine residues can only be mono- or di-methylated. S-adenosylmethionine (SAM) is used as the methyl donor for both lysine and arginine. Histone lysine methyltransferases (HKMT's) are responsible for the addition of a methyl group to lysines, and are specific for a particular residue in the histone tail [115]. Protein arginine methyltransferases (PRMT's) [67] are responsible for the addition of a methyl group to arginine residues. They come in two forms; the type I and the type II enzymes[115]. As with CpG methylation, histone methylation is also a dynamic mark and methyl groups can be removed from the histone.

Acetylation is another common histone modification. Acetyl groups can be added to lysine residues through the activity of histone acetlytransferases (HATs), which use acetyl Co-A as a donor. There are two major classes of HATs, those acetylating free histones (type-B HATs), and those acetylating bound histones (type-A HATs) [115]. The addition of an acetyl group to lysine neutralises its positive charge, and can weaken the interaction between the histone and the DNA, allowing transcription [67]. Conversely, histone deacetylases (HDACs) remove acetyl groups from lysine residues, stabilising the interaction between the histone and the DNA, and repressing

transcription. There are four classes of HDACs. Class I and II HDACs are most closely related to yeast scRpd3 and scHdaI respectively, class III HDACs are homologous to yeast scSir2, and there is only one class IV HDAC, HDAC11 [115].

Phosphorylation can occur on serine, threonine and tyrosine residues of a histone tail. This mark is also dynamic, with kinases transferring a phosphate group from ATP [115] onto the histone tail, and phosphatases removing this group [67]. The addition of a negative charge to the histone could act to further weaken the interaction between the histone and the DNA, promoting transcription through the region [114].

Other histone modifications include deimination (the conversion of an arginine to a citrulline), β-N-acetylglucosamine (the addition of β-N-acetylglucosamine to a serine or threonine residue), ADP ribosylation (both mono- and poly- ADP ribosylation of gluamtate and arginine residues), ubiquitylation (on lysine residues) and sumoylation (on lysine residues). All of these marks are reversible [115].

1.3.4.2. – Reading the Histone Code

The histone code is complex and dynamic. 'Writer' proteins place the modifications on the histone tails (HKMT's, PRMT's, HAT's and kinases), while 'eraser's' remove these modifications (HDAC's, and phosphatases). A third group of proteins, the 'readers', interpret these modifications to influence transcription through the underlying DNA. Each 'reader' tends to recognise specific modifications. In general methylated residues are recognized by proteins containing chromodomains, tudor domains, PHD fingers, MBT domains, Ankyrin repeats, PWWP domains, HEAT domains or WD40 domains [116]. Acetylated residues are recognized by proteins containing bromodomains, a

small domain containing a 4-helix-bundle fold and a hydrophobic binding pocket [116] [117]. Phosphorylated resiudes are recognized by proteins containing 14-3-3, BRCT, and BIR domains [116].

Histone modifications play a role in regulating gene transcription. Individual modifications tend to be associated with either activating or repressing transcription. As multiple modifications can occur on the same nucleosome it is possible for a bivalent state to occur, where the region contains both activating and repressive marks and is considered to be poised for future activation [110].

Examples of common histone modifications and their functions are listed in the table below (Table 1.3).

Modification	Function	Reference
H2A.Zub1	Repression or poising.	Santoro, 2015 [110]
acH2A.Z	Activation or poising.	Santoro, 2015 [110]
H2BK5ac	Activation.	Santoro, 2015 [110]
H2BK5me1	Elongation of transcription.	Santoro, 2015 [110]
H3K9me2	Repression.	Campos, 2009 [118], Zhou, 2011 [119]
H3K9me3	Repression.	Santoro, 2015 [110], Zhou, 2011 [119]
H3K4me3	Activation or poising.	Campos, 2009 [118], Santoro, 2015 [110], Zhou, 2011 [119]
H3K27me3	Repression or poising. It is associated with Polycomb repression.	Campos, 2009 [118], Santoro, 2015 [110], Zhou, 2011 [119]
H3K27ac	Activation.	Zhou, 2011 [119]
H3K36me3	Activation (when located in gene bodies).	Campos, 2009 [118], Zhou, 2011 [119]
H3K79me2	Activation (when located in gene bodies).	Zhou, 2011 [119]
H4K20me3	Repression.	Campos, 2009 [118]

Table 1.3 – Table of common histone modifications and their associated functions. Nomenclature of histone modifications – the first part specifies which histone the modification is on, the middle part specifies the location on the histone tail, and the final part specifies the type of modification. For example H3K9me3 means that lysine 9 on histone H3 is tri-methylated.

1.3.5. – Chromatin Architecture

The bound state of chromatin can affect gene expression. The bound state of a given region of DNA can vary by cell type and by phase of development. Generally, actively expressed genes are in regions of euchromatin, and repressed genes are in regions of heterochromatin. Euchromatin is less tightly coiled than heterochromatin, making the DNA more accessible for the transcriptional machinery [120].

Regions of DNA can interact with each other across quite large areas; for example the enhancer of Sonic Hedgehog (SHH) is located 1Mb away from the gene [121]. This

interaction occurs through looping of the DNA, which allows these two regions to localise together so they can interact. These loops are dynamic [122], and are usually associated with cohesin [123], which acts to hold both regions of DNA together to 'close' the loop, bringing the genes and regulatory elements located within it sufficiently close to interact.

CTCF recruits cohesin to its binding sites by interacting directly with its SA2 subunit. This stabilises long-range interactions between CTCF sites [84]. Although CTCF and cohesin work together to form chromatin loops to regulate long-range chromatin interactions, they play different roles. Experiments in which their binding was disrupted in human cell lines showed that both cohesin and CTCF are responsible for local chromatin interactions, but CTCF also plays a role in silencing interactions between different loops [124].

Cohesin has also been shown to interact with the Mediator Complex (see section 1.3.6.) to form loops to regulate the expression of lineage specific genes [125]. The Mediator Complex interacts with cohesin through NIPBL, which is involved in the loading of cohesin onto DNA [65].

The development of Hi-C, which allows the interactions between all genomic regions to be determined [126], has shown that the genome is portioned into discrete compartments approximately 500-900Kb in length [122]. These are known as topologically associated domains (TADs), and they are conserved across mammals and cell types. Loci within a TAD more frequently form contacts with each other than they do with loci in other TADs [127]. Cohesin and CTCF are enriched at TAD borders

[126] [128], which could help to explain why most interactions are within TADs rather than across them. The genes within a TAD tend to be co-ordinately regulated. CTCF mediated chromatin loops have also been implicated in maintaining TAD structure [128].

1.3.6. – RNA polymerase II and the Mediator Complex

A variety of proteins can bind to the gene or its regulatory regions to mediate its expression. TFs bind to the promoter or enhancer of a gene and recruit RNA polymerase II to the TSS, through the Mediator Complex. The mammalian Mediator Complex consists of 26 subunits, which can be exchanged, making it a dynamic complex. Different TFs bind to different subunits, allowing for the binding of multiple TFs at the same time [64]. The Mediator Complex then recruits RNA polymerase II through an interaction between the C-terminal domain (CTD) of the RPBI subunit (the largest subunit) of RNA polymerase II [64] [65]. This forms the basis of the Pre-Initiation Complex (PIC) along with TFIIA, TFIIB, TFIID, TFIIE, TFIIF and TFIIH [125], and transcription is initiated.

At many genes RNA polymerase pauses in its transcription between 30 and 50 nucleotides downstream of the TSS [65]. The Mediator Complex plays a role in this pausing, but the mechanism through which it acts is still unclear [64]. RNA polymerase pausing can also be due to the nucleosome occupancy of the chromatin, or the methylation state of the DNA. This pausing can affect exon and intron inclusion in the transcript, which will be discussed further in section 1.3.8. Differential inclusion of exons is just one of the ways of generating such a large diversity of transcripts (about 200, 000) from the relatively small number of genes (about 20, 000) found in humans.

1.3.7. – Alternative Promoters

One way to increase transcript diversity from genes is through the use of alternative promoters, producing transcripts that vary in their 5' end. For example *Grb10* (growth-factor receptor bound protein 10), an imprinted signal adaptor protein contains four promoters. *Grb10* is expressed from the maternal methylated allele in most mouse tissues except for a subset of neurons, where it is expressed from the un-methylated paternal allele. Maternal transcripts always originate from the major promoter of the gene. Paternal transcripts originate from one of three minor promoters located downstream of the major promoter [129]. In this example the use of alternative promoters is being used to control tissue-specific expression of the gene.

1.3.8. – Alternative Splicing

It is estimated that around 95% of multiexonic human genes are subject to alternative splicing of the pre-mRNA [130] making this an important mechanism to consider when investigating the regulation of gene expression. Alternative splicing is the mechanism through which multiple different mRNA transcripts can be generated from a single gene [131]. There are five types of alternative splicing; exon skipping, the use of alternative 5' splice sites, the use of alternative 3' slice sites, intron retention and mutually exclusive splicing. Exon skipping is responsible for 38% of alternative splicing events conserved between mice and humans. The use of alternative 5' and 3' splice sites account for 18% and 8% respectively. In this case all exons are represented in the RNA transcript but only part of the sequence may be retained for one or more of them depending on which 5' or 3' splice site in the exon is used. Intron retention accounts for 3% of splicing events and occurs when an intron that is normally removed from the

transcript is retained. The final 33% of splicing events are more complex cases, and include mutually exclusive exon inclusion, where the presence of one exon in the transcript inhibits the inclusion of another [132].

Pausing of the RNA polymerase can lead to alternative splicing. RNA polymerase pausing can be caused by nucleosome occupancy of the chromatin, or by methylation of the DNA sequence. Exons tend to have a higher nucleosome occupancy than introns, and both exons and introns have specific histone modifications. RNA polymerase pauses before histones at exons, regulating the inclusion rate of particular exons in the RNA strand. Splicing factors can be recruited to the growing RNA strand while transcription is occurring, and a slowing of the rate of transcription can give these factors more time to recognise splice sites [133]. On average, exons have a higher level of DNA methylation than introns, which suggests that methylation status could influence alternative splicing [131]. CTCF and MeCP2 are both involved in translating these methylation marks. At the *CD45* locus, CTCF binding at the un-methylated boundary between the intron and the alternatively spliced exon (ASE) (exon 5) leads to RNA polymerase pausing and inclusion of exon 5 into the transcribing RNA [134]. MeCP2 has been found to bind to methylated boundaries between introns and ASEs, leading to the inclusion of the ASE. This is thought to be due to both the pausing of the RNA polymerase caused by the presence of MeCP2 on the DNA, and the recruitment of HDACs to deactylate the histones in the associated ASE (by MeCP2) [135].

The regulation of alternative splicing is complicated and involves many different factors. Further work is needed to investigate how all these factors work together to ensure the correct splicing of the gene.

1.3.9. - Alternative Polyadenylation

All fully processed eukaryotic mRNA's have a 3' poly (A) tail, a region of the mRNA consisting only of adenines [66]. The addition of a poly (A) tail is specified by three sequence elements (Figure 1.11); the upstream sequence element (USE), consensus polyadenylation signal (PAS) and the downstream sequence element (DSE) [136]. Although these sequences are traditionally associated with the 3' end of the gene, they can be located throughout the gene body, and each gene can contain multiple polyadenylation sites. About 54% of human genes and 32% of mouse genes take advantage of these alternative polyadenylation sites to generate different transcripts from the same gene [137].

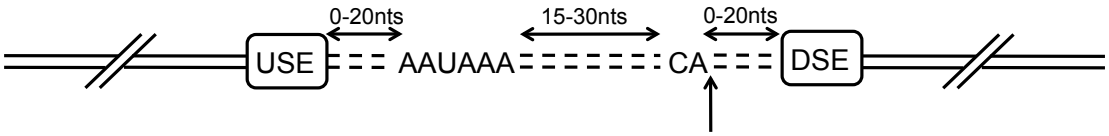


Figure 1.11 – Diagram of the consensus sequence elements required for polyadenylation in mammals. USE = the U-rich upstream sequence element. DSE = the G/U- or U-rich downstream sequence element. nts = nucleotides. The arrow shows where cleavage occurs. Adapted from (Proudfoot, 2011 [136]).

The polyadenylation site consists of a consensus polyadenylation signal (PAS) located between two Uracil-rich regions. The PAS consensus sequence is AAUAAA, but this motif can vary slightly. It is located between 0-20 nucleotides downstream of the USE. 15-30 nucleotides downstream of the PAS is a CA dinucleotide, where cleavage of the RNA strand occurs. 0-20 nucleotides downstream of this is the G/U- or U-region known as the DSE. Cleavage and polyadenylation specificity factor (CPSF) and cleavage stimulatory factor (CstF) recognise the AAUAAA and DSE and act to cleave the RNA strand between these two elements at the CA dinucleotide [136].

1.3.10. – Transcriptional Interference

Transcriptional interference occurs when one transcriptional process or RNA polymerase complex has a repressive effect on a second transcriptional process [138]. This usually occurs between two genes where their promoters are either convergent, located in tandem or are in an overlapping divergent configuration (Figure 1.12). There are five mechanisms through which this can occur; promoter competition, the sitting duck mechanism, occlusion, collision and road block.

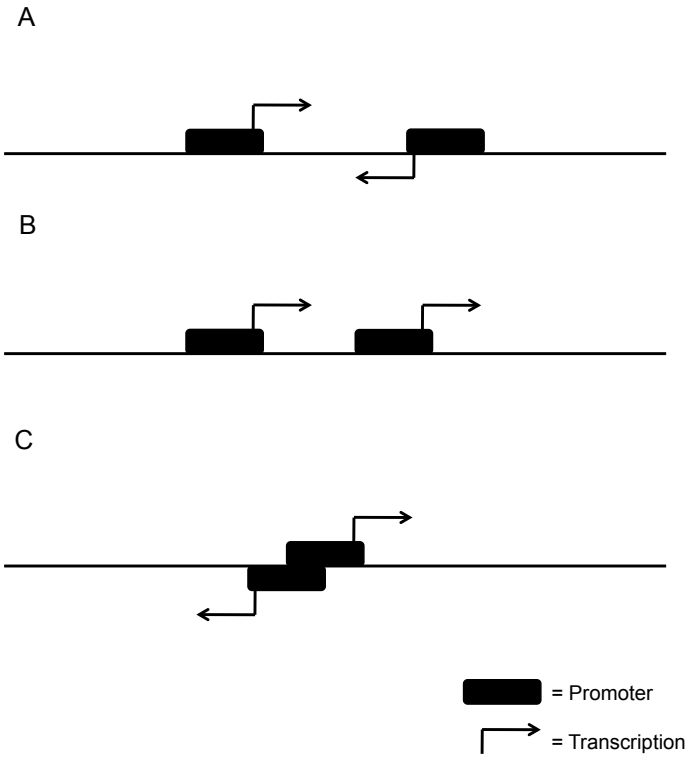


Figure 1.12 – Different promoter arrangements important for transcriptional interference. A – Convergent promoters. B – Tandem promoters. C – Overlapping divergent promoters. Adapted from (Shearwin, 2005 [139]).

Promoter competition occurs where the binding of an RNA polymerase complex to one promoter inhibits RNA polymerase binding at the second promoter (Figure 1.13 A). This can occur at all three of the promoter arrangements.

The sitting duck mechanism occurs where one of the two promoters is 'stronger' (quicker to transition from the open RNA polymerase complex to the more closely associated elongating RNA polymerase complex) than the other. The RNA polymerase complex originating from the 'strong' promoter can hit and knock the transcription-initiating complex from the second 'slower' promoter (Figure 1.13 B). This can occur at both convergent and tandem promoters.

Occlusion occurs where transcription from the first promoter across the second promoter transiently blocks transcription from the second promoter (Figure 1.13 C). This can occur at promoters located convergently or tandemly to each other.

Collisions between converging elongation complexes can result in one or both of the elongation complexes being dislodged from the DNA, leading to premature termination of the transcript (Figure 1.13 D).

So far the road block mechanism has only been seen where the DNA-bound Lac repressor acts to block the progress of RNA polymerase initiating upstream of the bound Lac repressor (Figure 1.13 E). In theory this could occur where an open RNA polymerase is so tightly bound to the promoter that it can act as an immovable road block rather than a sitting duck, but this has not been confirmed [139]. These mechanisms can all cause premature termination of one or both of the transcripts involved.

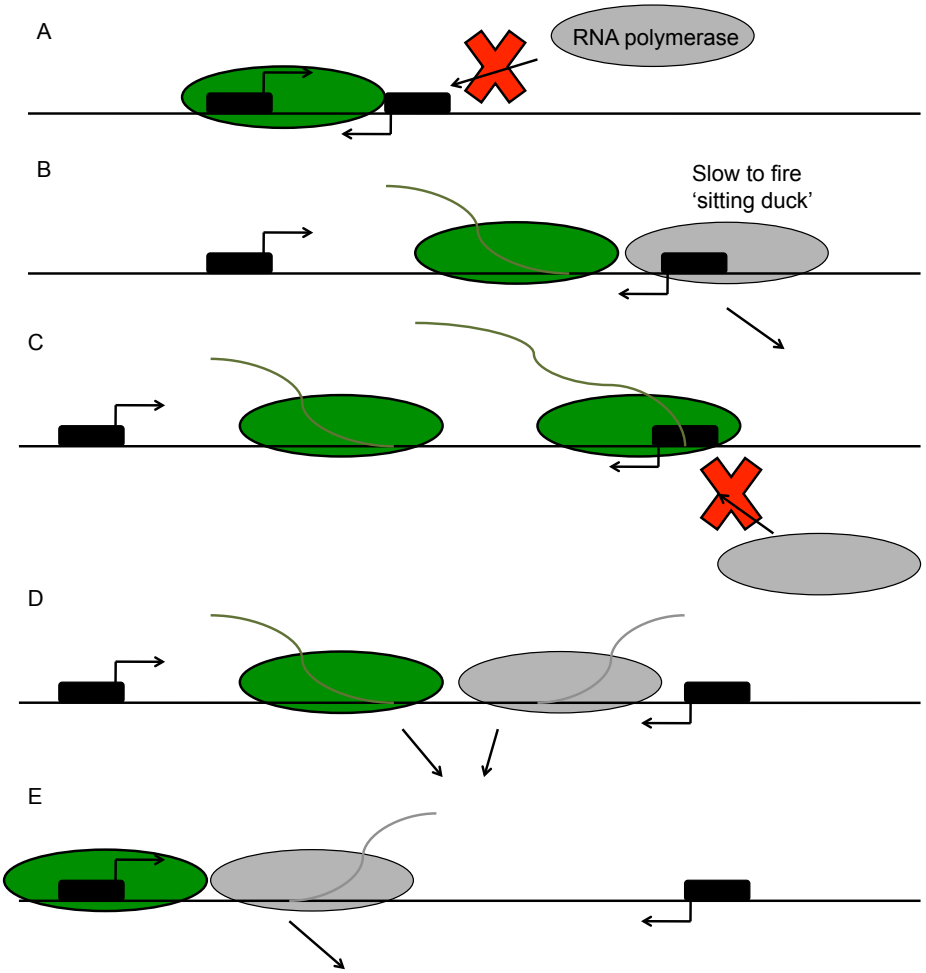


Figure 1.13 – Mechanisms of transcriptional interference. A – Promoter competition. B – Sitting duck mechanism. C – Occlusion. D – Collision. E – Road block. Adapted from (Shearwin, 2005 [139]).

Some promoters are capable of initiating transcription in both the sense and antisense directions, ie they are bidirectional [140] [141]. Initiation rates have been shown to be roughly equal for both orientations, but transcription in the antisense orientation tends to drop off as RNA polymerase II moves further away from the promoter resulting in the generation of more sense than antisense transcripts [141]. However there are still a large number of genes where the antisense transcript plays an important role, for example *Tsix*, the antisense transcript of *Xist*, protects the active X chromosome from inactivation during random X chromosome inactivation [8]. The bidirectional nature of

promoters means that several mechanisms of transcriptional interference may be acting on any promoter pair at a given time.

Imprinted retrogenes are an interesting case for investigating these models of transcriptional interference. Retrogenes are generated from the retrotransposition of an mRNA molecule into the genome [142]. In some cases the mRNA molecule is incorporated into an intron of a 'host' gene, and it is this type of retrogene that we are interested in studying. They tend to be monoexonic and retain their function and activity [143]. Imprinted retrogenes (which are situated in the intron of a non-imprinted gene) are good models for examining gene regulation; although the two alleles share an identical environment and are subject to the same influences, the retrogene is only expressed from one allele and not the other. Therefore, this must be due to epigenetic factors operating *in cis*. The methylated CpG dinucleotides at the promoter of the repressed allele act to both block the binding of transcription factors and recruit methylation binding proteins, which also act to inhibit the binding of transcription factors to the gene [42]. CpG methylation can also act to alter the density of the chromatin over the region in which it occurs, through the activity of histone modifying enzymes. Dnmt1 and Dnmt3a have been shown to bind to a histone methyltransferase, and Dnmt1 and Dnmt3b can bind to histone deactetylases, recruiting them to regions of DNA methylation. The addition of methyl groups and the removal of acetyl groups at the histone tails increases the binding affinity of the histone for the DNA, leading to tighter binding blocking access of transcription factors to the gene [42]. MeCP2 can also recruit histone deacetylases to regions of methylated CpG dinucleotides to the same effect [42], and has been shown to recruit histone methyltransferases to reinforce the repression of the gene [144].

DNA methylation is not constant across a locus, not all of the CpG dinucleotides in a region classified as methylated are actually methylated, and the ones that actually are can vary between cells and tissue types. The imprinted locus *Gnas*, which includes the imprinted transcripts *Gnas, Gnasxl*, *Exon 1A*, *Nesp* and *Nespas*, is an example of a region where the methylation status varies across the locus. It contains three DMR's, one which is unmethylated on the maternal allele and methylated on the paternal allele, and two which are methylated on the maternal allele and unmethylated on the paternal allele [145].

The work presented in this thesis makes use of the *H13/Mcts2* locus as a model of a host/imprinted retrogene pair.

1.4. – The *H13/Mcts2* locus

Mcts2 is an imprinted retrogene, located within intron four of *H13*, its host gene, on chromosome 2. The promoter of *Mcts2* is methylated on the maternal allele, and so *Mcts2* is silenced, and *H13* is transcribed using a down-stream poly (A) site. However on the paternal allele, the promoter is un-methylated and *Mcts2* is expressed. The expression of *Mcts2* coincides with the premature termination of the *H13* transcript (see Figure 1.14), through the use of an alternative upstream poly (A) site.

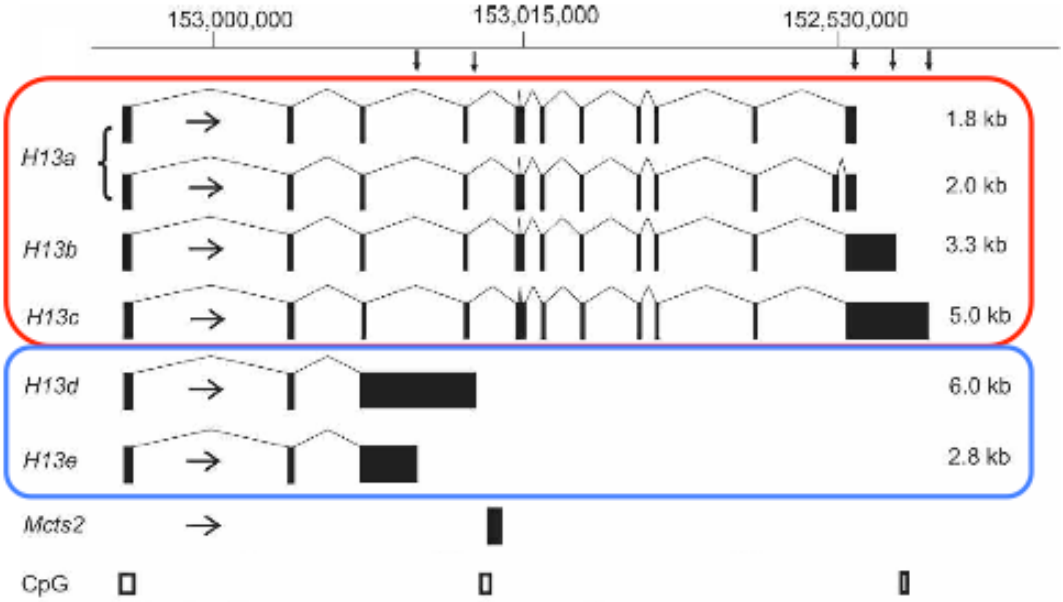


Figure 1.14 – A summary of the transcripts generated from the *H13* locus. The black boxes represent the exons included in each transcript, while the lines joining them represent the introns excluded. In the case of *H13d* and *e* parts of intron 3 are retained in the transcript. The transcripts generated from the maternal allele are shown in the red box, and the transcripts generated from the paternal allele are shown in the blue box. The location of *Mcts2* in intron 4 is shown, as are the locations of the CpG islands associated with this locus (reproduced from Wood, 2008 [146]).

There are several internal poly (A) sites in the *H13* locus, generating at least five different transcripts. There are three transcripts (*H13a*, *b* and *c*) generated from the maternal allele utilising downstream poly (A) sites, and two transcripts (*H13d* and *e*) generated from the paternal allele utilising up-stream poly (A) sites (Figure 1.14) [146]. The fact that *H13d* and *e* are only generated when *Mcts2* is expressed suggests that transcription from an internal site could be responsible for premature termination of the host transcript. It is unclear whether this premature termination is caused by the transcription of the retrogene or by the binding of a methylation sensitive polyadenylation factor to the DMR. For example, this could be a factor that binds in the presence of methylation and inhibits the use of upstream poly (A) sites, dissociating in the absence of methylation to allow the use of both upstream and downstream poly (A) sites (with upstream sites being used preferentially in the case of *H13*). Alternatively,

such a factor could bind only in the absence of methylation, promoting the use of upstream poly (A) sites, and therefore when the DMR is methylated there is no binding and only downstream poly (A) sites are used.

1.5. – Project Aims

The aim of this project is to further our understanding of the mechanisms of gene regulation at imprinted loci, using a model locus (the *H13/Mcts2* locus) to explore some of these mechanisms in more detail. This is being investigated using a variety of different experimental and bioinformatics approaches. Experimental approaches are being used to demonstrate that intragenic promoters affect host gene transcript polyadenylation; through the deletion of the intragenic promoter and the relocation of the intragenic promoter into a new host gene. The significance of epigenetic mechanisms in alternative polyadenylation and premature termination of a transcript is being determined through the use of siRNA's to alter histone methylation marks, and to investigate if this mechanism occurs at other loci in the genome, in addition to the *H13/Mcts2* locus. Experimental and bioinformatics approaches are being used to discriminate between the models of transcriptional interference by an internal promoter and differential binding across the CpG island of the internal promoter. The work in this thesis investigates the affect of intragenic promoters on host gene transcription, and aims to discriminate between the models of transcriptional interference and differential binding across the CpG island of the internal promoter.

I was involved in the previous studies from Professor Oakey's laboratory, which used a ChIP-Seq approach to map genomewide the locations of CTCF and cohesin binding [22]. ATRX and MeCP2 have been shown to bind with CTCF and cohesin at two

individual loci in the genome, the *H19* ICR and the *Gtl2/Dlk1* imprinted region [147]. In chapter three I will describe the use of ChIP-Seq to identify the binding sites of ATRX and MeCP2. By combining these with the existing data for CTCF and cohesin it will be possible to investigate the interaction of these four proteins in a genomewide manner. Identifying regions which co-bind these proteins will help to generate a model to inform our understanding of how these proteins influence gene expression, particularly at other imprinted loci located across the genome.

In chapter four I will focus on an individual locus, *H13/Mcts2*, and investigate its specific mechanisms of gene regulation. Expression of *Mcts2* is associated with altered transcription through *H13*, where the use of poly (A) sites upstream of *Mcts2*, leads to premature termination of the *H13* transcript. We hypothesise that this premature termination is caused by transcription of the retrogene interfering with host gene transcription. I will describe the design of two series of DNA constructs based on this locus. The first will allow expression through the retrogene to be regulated, allowing the direct consequences of *Mcts2* expression to be examined in detail. The second inserts *Mcts2* into *Fam13c*, a large poly-exonic gene similar in structure to *H13*. This will enable us to investigate whether the presence of *Mcts2* can enforce the same phenotype, premature transcript polyadenylation, on *Fam13c* as it does at *H13*. These studies will provide a mechanistic component to our whole genome analyses, allowing us to better understand the role of DMRs and propose specific mechanisms through which gene expression is epigenetically regulated.

Chapter 2

Materials and Methods

2.1. - Source of Mouse Tissue

Tissue was generated from crosses between mice from two inbred subspecies: *Mus musculus castaneus* and C57BL/6 mice. These mice were housed at the Biological Services Unit at King's College London according to the United Kingdom Home Office Breeding Licence. Colonies were managed with help from Dr Adam Prickett, Dr Michael Cowley and Matthew Shannon. The mice were fed on expanded breeder pellet R&M No 3 (Special Diets Services, Essex, UK), and kept in a 12 hour dark/light cycle with 30 minutes dawn/dusk lighting, an ambient temperature of 21°C ± 2°C and at a humidity of 50% ±10%.

2.2. - Cell Culture

Cell lines were sourced from the American Type Culture Collection (ATCC). All cell lines used were cultured at 37°C in a 5% CO_2 atmosphere. All reagents were heated to 37°C before use. Cells were cultured in tissue culture rated vessels and passaged when they reached 80% confluency, as determined by visual inspection. To passage the 3T3's the medium was aspirated, and the cells were washed twice with PBS. The cells were incubated in 0.05% trypsin in EDTA (Gibco®, Life Technologies™, cat number 25300-054) at 37°C until they detached from the flask. 10mls of medium was added to the flask to inactivate the trypsin. The cells were transferred via 10ml stripette to a 50ml falcon tube, and centrifuged at 1,000RPM for 5 minutes at room temperature. The medium was removed and the cells were resuspended in fresh medium and transferred to a new flask. To passage the Neuro2a and HEK 293 cells the medium was removed

and replaced with 10mls of fresh medium. The medium was washed over the back of the flask (over the surface where the cells have attached) to detach them from the flask. The detached cells and medium were transferred to a 50ml falcon tube, and centrifuged at 1,000RPM for 5 minutes at room temperature. The medium was removed and the cells were resuspended in fresh medium and transferred to a new flask. All three of the cell lines were split at a range of concentrations from 1 in 10 to 1 in 20, depending on their growth rate.

BxJ and JxB Embryonic Stem (ES) cells were kindly donated by Dr Robert Feil (Institut de Génétique Moléculaire de Montpellier). These cells were grown on 0.1% gelatin coated plates. Plates were coated with an excess of ESGRO Complete Gelatin Solution (Millipore, cat number SF008) for at least 30 minutes at room temperature. The excess was removed. The medium (Merck Millipore, cat number SF001-500P) was aliquoted into 50ml falcon tubes, to which 25µl of GSK3 β Inhibitor and 5 µl of LIF (Millipore, cat number ESG1106) was added just before use.

Cell lines			BxJ and JxB ES cells
Neuro2a	3T3	HEK 293	
Dulbecco's Modified Eagle Medium (DMEM)	Dulbecco's Modified Eagle Medium (DMEM)	Dulbecco's Modified Eagle Medium (DMEM)	ESGRO Complete PLUS Clonal Grade Medium
10% Fetal Bovine Serum (FBS)	10% Fetal Bovine Serum (FBS)	10% Fetal Bovine Serum (FBS)	GSK3 β Inhibitor
Penicillin (at a final concentration of 1 unit/ml)	Penicillin (at a final concentration of 1 unit/ml)	Penicillin (at a final concentration of 1 unit/ml)	LIF (at a final concentration of 100 units/ml)
Streptomycin (at a final concentration of 1 ug/ml)	Streptomycin (at a final concentration of 1 ug/ml)	Streptomycin (at a final concentration of 1 ug/ml)	
MEM non-essential amino acids 1x			

Table 2.1 – Table detailing the type of medium used for each cell line. FBS was filter sterilised before being added to the medium. All the components of the medium for the cell lines were from Life Technologies. Gibco® MEM Non-Essential Amino Acids (cat number 11140-035). Penicillin and streptomycin (cat number 15140-122). DMEM. The medium for the ES cells was from Millipore. ESGRO Complete PLUS Clonal Grade Medium (cat number SF001-500P). The GSK3 β Inhibitor was supplied with this medium.

2.3. – Chapter 3

2.3.1. – Chromatin Extraction from Tissues

This protocol was used to determine the allele-specific binding of CTCF and cohesin (as represented by Rad21) in postnatal day 21 mouse brain tissue [22].

Chromatin was extracted from p21 BxC and CxB inter-cross mouse brain tissue. The tissue was first cut into small pieces with a sterile scalpel and then homogenized in 1ml of PBS pH8, using a Dounce homogeniser (Fisher). The homogenised tissue was then centrifuged at 7,000RPM for 5 minutes at 4°C. The supernatant was removed, the

nuclei were resuspended in 1ml PBS and then subjected to centrifugation at 7,000RPM for 3 minutes at 4°C. The nuclei were washed in PBS three more times. After the fourth wash the nuclei were resuspended in 1ml 5mM DTBP to cross-link the proteins to the chromatin. The nuclei were incubated on ice for 30 minutes, before being washed twice in PBS (as previously). The reaction was stopped by adding 1ml quench buffer (1M Tris HCL pH8, 5M sodium chloride and H_2O) to the nuclei. The nuclei were then washed twice in PBS (as previously). The nuclei were then cross-linked with 1% formaldehyde in PBS, and incubated on ice for 10 minutes. The nuclei were washed three times in PBS (as previously but for 4 minutes instead of 3). The nuclei were resuspended in 1ml lysis buffer (0.1M PMSF, Protease inhibitor, 1M Tris HCl pH8, 10% SDS, 0.5M EDTA, H_20), and then sonicated to break up the chromatin into roughly 500bp fragments (at 40AMP, 1 minute on, 1 minute off, repeated for 15 minutes). A small sample of the chromatin was incubated overnight at 65°C with 5M sodium chloride and H_20 to de-cross-link it so that it could be run on a gel to check the efficiency of the sonication. The concentration of the rest of the chromatin was measured and it was stored at -80°C.

2.3.2. – Chromatin Immunoprecipitation - Method 1

Chromatin (to a final concentration of 0.08µg/µl) was added to 80µl agarose beads (Millipore, cat number 16-156 for Protein A beads and cat number 16-266 for Protein G beads), EDTA-free protease inhibitor (to a final concentration of 1x) (Roche, cat number 04693132001) and dilution buffer. Samples were rotated at 4°C for 2 hrs. Each sample was applied to a Corning® Costar® Spin-X® centrifuge tube (Sigma Aldrich, cat number CLS8160-96EA) and centrifuged at 7,000g at 4°C for 2 minutes. This centrifugation step was repeated as needed, until the entire sample passed through

the column. 200µl was aliquoted into each tube. 140µl of dilution buffer was added to the tubes for the antibodies, and 200µl was added to the tube for the input sample. The input was then stored at 4°C until the phenol extraction stage later in the protocol. The antibodies were added to the appropriate tubes (20µg of ATRX; Insight Biotechnology, cat number sc-15408, 4µg of IgG; Millipore, cat number 06-371, and 7µg of MeCP2; ABCAM, cat number AB2828). The samples were rotated at 4°C overnight.

20µg of yeast tRNA, and 60µl agarose beads were added to each sample, and the samples rotated at 4°C for 2 hours. The samples were centrifuged at 7,000g at 4°C for 2 minutes, before being transferred to Corning® Costar® Spin-X® centrifuge tubes. The samples were centrifuged at 7,000g at 4°C for 2 minutes, and the flow-through discarded. 600µl of wash buffer 1 was added to the beads (in the same centrifuge tubes), and the samples were rotated at 4°C for 10 minutes. The samples were centrifuged at 7,000g at 4°C for 2 minutes, and the flow-through discarded. 600µl of wash buffer 2 was added to the beads (in the same centrifuge tubes), and the samples were rotated at 4°C for 10 minutes. The samples were centrifuged at 7,000g at 4°C for 2 minutes, and the flow-through discarded. 600µl of wash buffer 3 was added to the beads (in the same centrifuge tubes), and the samples were rotated at 4°C for 10 minutes. The samples were centrifuged at 7,000g at 4°C for 2 minutes, and the flow-through discarded. The beads were resuspended in 400µl of ddH_2O and moved to a 2ml microcentrifuge tube. The input sample was processed from this point on. 16µl of 5M sodium chloride was added to each sample, and incubated at 65°C overnight.

2µl proteinase K (Roche, cat number 03115887001) and 1µl RNAse A was added to each sample and incubated at 45°C for 2 hrs. Phase lock tubes (VWR, cat number 713-

2533) were prepared by centrifuging them at 13,000g at 4°C for 1 minute. One volume of phenol:chloroform was added to the DNA in a phase lock tube and inverted to mix the solution. The solution was then centrifuged at 13,000RPM at 4°C for 10 minutes. The top aqueous layer was removed and transferred to a fresh microfuge tube. $1/10^{th}$ volume of 5M sodium chloride, three times the volume of 100% ethanol and 1ul of glycogen were added to each tube. The sample was then incubated at -20°C for 30 minutes before being centrifuged at 13,000RPM at 4°C for 20 minutes. The supernatant was removed and the pellet was washed in 70% ethanol. The tube was centrifuged at 13,000RPM at 4°C for 10 minutes, and the ethanol wash was then removed. The pellet was left to dry before the DNA was resuspended in 21µl ddH_20. The concentration was measured on the Qubit®, before being stored at 4°C.

2.3.3. - Chromatin Immunoprecipitation - Method 2

The EZ-ChIP™ kit (Millipore, cat number 17-371).

4.5µl of Protease Inhibitor Cocktail II was added to 900µl dilution buffer (for each IP) and stored on ice. 900µl of the prepared dilution buffer was added to 100µl of sheared, cross-linked chromatin. 60µl of protein G agarose beads were added, and the samples were incubated for 1 hour at 4°C, rotating. The samples were centrifuged at 5,000g for 1 minute to pellet the beads. 10µl of the supernatant was removed and saved as the Input at 4°C until the elution step on the following day. The remaining supernatant was moved to a fresh microfuge tube, and the appropriate antibodies were added (20µg of ATRX; Insight Biotechnology, cat number sc-15408, 4µg of IgG; Millipore, cat number 06-371, and 7µg of MeCP2; Abcam, cat number ab2828). Samples were incubated overnight at 4°C.

60µl of protein G agarose beads were added to each sample, and incubated for 1 hour at 4°C, rotating. Samples were centrifuged at 5,000g for 1 minute and the supernatant was removed. The beads were resuspended in 1ml of cold Low Salt Immune Complex Wash Buffer, and incubated for 5 minutes, rotating. Samples were centrifuged at 5,000g for 1 minute and the supernatant was removed. The beads were resuspended in 1ml of cold High Salt Immune Complex Wash Buffer, and incubated for 5 minutes, rotating. Samples were centrifuged at 5,000g for 1 minute and the supernatant was removed. The beads were resuspended in 1ml of cold LiCl Immune Complex Wash Buffer, and incubated for 5 minutes, rotating. Samples were centrifuged at 5,000g for 1 minute and the supernatant was removed. The beads were resuspended in 1ml of cold TE Buffer, and incubated for 5 minutes, rotating. Samples were centrifuged at 5,000g for 1 minute and the supernatant was removed. This wash in TE buffer was repeated again.

Elution of protein/DNA complexes:

For each tube 200µl of elution buffer was prepared; 10µl 20% SDS, 20µl 1M $NaHCO_3$ (warmed to room temperature) and 170µl sterile, distilled H_2O. 200µl was added to the Input and left at room temperature. For sample tubes 100µl elution buffer was added and mixed by flicking. Sample tubes were incubated for 15 minutes at room temperature, and centrifuged at 5,000g for 1 minute. The supernatant was moved to a fresh tube. 100µl of elution buffer was added to the beads, and the beads were resuspended by flicking the tube. The tubes were incubated for 15 minutes at room temperature, and centrifuged at 5,000g for 1 minute. The supernatant was pooled with the previous eluted DNA.

Reverse the cross-links of the protein/DNA complexes to free the DNA:

8µl of 5M NaCl was added to each tube, and samples were incubated overnight at 65°C. 1µl of RNase A was added to each tube, and samples were incubated for 30 minutes at 37°C. 4µl of 0.5M EDTA, 8µl 1M Tris-HCl and 1µl Proteinase K was added to each tube, and incubated for 1-2 hours at 45°C.

DNA purification using spin columns:

1ml of Bind Reagent A was added to each 200µl DNA sample tube, and mixed well. 600µl of sample/Bind Reagent A was transferred to a spin filter in a collection tube, and centrifuged at 10, 000g for 30 seconds. The liquid was discarded and the sample/Bind Reagent A was again centrifuged at 10, 000g for 30 seconds, and the liquid discarded. 500µl of Wash Reagent B was added to the spin filter, and centrifuged at 10, 000g for 30 seconds. The liquid was discarded and the spin filter was centrifuged at 10, 000g for 30 seconds. The spin filter was transferred to a fresh collection tube and 50µl of Elution Buffer C was added to the centre of the spin filter membrane. The spin filter was centrifuged at 10, 000g for 30 seconds. The eluted DNA was stored at -20°C.

2.3.4. - Chromatin Immunoprecipitation – Method 3

To prepare the Dynabeads/antibody:

50µl of Dynabeads Protein A magnetic beads (Life Technologies™, 10001D) were aliquoted into each tube and 1ml of filter sterilised 5mg/ml PBS/BSA was added. The tubes were placed in a magnetic rack (Ambion®, AM10055), and the PBS/BSA solution was removed. The beads were washed in this way three more times. The beads were resuspended in the appropriate antibodies (20µg of ATRX; Insight Biotechnology, cat number sc-15408, 4µg of IgG; Millipore, cat number 06-371, and

7µg of MeCP2; Abcam, cat number ab2828) and up to 300µl of PBS/BSA was added. The tubes were incubated, rotating, at 4°C for 3-4 hours. The beads were then washed four times in 1ml of PBS/BSA, before being resuspended in 300µl of PBS/BSA.

To prepare the Dynabeads for pre-clearing:

50µl of Dynabeads Protein A magnetic beads (Life Technologies™, 10001D) were aliquoted into each tube and 1ml of filter sterilised 5mg/ml PBS/BSA was added. The tubes were placed in a magnetic rack, and the PBS/BSA solution was removed. The beads were washed in this way three more times. The beads were resuspended in 300ul of PBS/BSA and incubated, rotating, at 4°C for 3-4 hours.

To prepare the chromatin:

Neuro2a cells/ JxB and BxJ ES cells were incubated with 0.05% trypsin (Gibco®, Life Technologies™, cat number 25300-054) at 37°C until they detached from the flask. 10mls of DMEM (with 10% Fetal Bovine Serum and Penicillin and Streptomycin) was added to the flask to inactivate the trypsin. The cells were transferred to a 50ml falcon tube, and centrifuged at 1,000RPM for 5 minutes. The medium was removed and 10mls of PBS was added. The cells were centrifuged at 1,000RPM for 5 minutes. The PBS wash was repeated once more. The cell pellet was resuspended in 30mls PBS and the cells were counted. 20 million cells were used for each ChIP reaction. EGS (Pierce, cat number 21565, made up fresh) was added to each reaction to a final concentration of 2mM, and the cells were incubated at room temperature for 45 minutes on a roller. Formaldehyde solution was added to each reaction to a final concentration of 1%, and the cells were incubated at room temperature for 20 minutes. 1/8 volume of 1M glycine

was added to each reaction to quench the formaldehyde. The cells were centrifuged at 1,000RPM for 5 minutes. The cells were washed twice with 20ml 4°C PBS.

The cells were resuspended in 10ml of lysis buffer 1 (see appendix) and rocked at 4°C for 10 minutes. They were centrifuged at 2,000RPM at 4°C for 2 minutes. The cells were resuspended in 10ml of lysis buffer 2 (see appendix) and rocked at 4°C for 10 minutes. They were centrifuged at 2,500RPM at 4°C for 2 minutes. The cells were resuspended in 3ml of lysis buffer 3 (see appendix). The cells were aliquoted into 1.5ml microfuge tubes (1ml into each tube). The chromatin was sonicated to break up the chromatin into roughly 500bp fragments (at 40AMP, 1 minute on, 1 minute off, repeated for 15 minutes). After sonication add 1/10 the volume of 10% Triton-X 100 to each tube, and centrifuged at 14,000 RPM at 4°C for 10 minutes. The lysate was transferred to a 15ml falcon tube and the volume adjusted to 3ml. 50µl of the lysate from each sample was saved overnight at -20°C as an input sample. 50µl was saved to assess the sonication efficiency.

300µl of the pre-clearing beads were added to 3ml of chromatin. The chromatin/bead mix was incubated, rotating, at 4°C for 1 hour. The samples were applied to a magnet and the chromatin was removed to fresh 5ml tubes. The washed antibody bound beads were added, and the samples were incubated, rotating, at 4°C overnight.

The beads were moved to a 1.5ml microfuge tube. 1ml of wash buffer was added and the samples incubated, rotating, at 4°C for 5 minutes. The samples were placed in a magnetic rack and the supernatant was removed. The wash was repeated four times. The samples were then washed once in 1ml cold PBS (with protease inhibitors). All of

the five washes in wash buffer and the PBS wash were carried out in the cold room (at 4°C) with ice-cold solutions. After the PBS was removed from the beads they were resuspended in 250µl elution buffer, and incubated, rotating, at room temperature for 15 minutes. The samples were placed in a magnetic rack and the supernatant was removed to a fresh 1.5ml microfuge tube. A further 250µl elution buffer was added to the beads, and they were incubated, rotating, at room temperature for 15 minutes. The samples were placed in a magnetic rack and the supernatant was added to that from the previous round of elution. 20ul of 5M sodium chloride was added to the samples and they were incubated at 65°C for 4 hours. The frozen input was thawed and 450µl of elution buffer, and 20µl 5M sodium chloride was added. The input samples were also incubated at 65°C for 4 hours.

20ul of 1M Tris pH6.5, 10ul of 0.5M EDTA, 2ul of 10mg/ml Proteinase K and 1ul of glycogen were added to each sample. The samples were then incubated at 45°C for 1 hour. A phenol extraction and ethanol precipitation was then performed to clean the extracted DNA. Phase lock tubes (VWR, cat number 713-2533) were prepared by centrifuging them at 13,000g at 4°C for 1 minute. One volume of phenol:chloroform was added to the DNA in a phase lock tube and inverted to mix the solution. The solution was then centrifuged at 13,000RPM at 4°C for 10 minutes. The top aqueous layer was removed and transferred to a fresh microfuge tube. 50ul of 5M sodium chloride, 1,500ul of 100% ethanol and 1ul of glycogen were added to each tube. The solution was then incubated at -20°C for 30 minutes before being centrifuged at 20,000g at 4°C for 10 minutes. The supernatant was removed and the pellet was washed in 70% ethanol. The tube was centrifuged at 20,000g at 4°C for 5 minutes, and the ethanol wash was then removed. The pellet was left to dry before the DNA was resuspended in

21µl ddH_20. The concentration was measured on the Qubit®, before being stored at 4°C.

2.3.5. - Quantitative Real-Time PCR

To check the efficiency of the ChIP qPCR was performed using Custom Plus TaqMan™ RNA Assays (Applied Biosystems) or primers used in conjunction with the Roche Universal Probe Library (Roche). For more details see appendix 7.2.

3µl of the ChIP sample was used as the template in a PCR reaction with 0.5µl primers and probe (at a 10mM concentration), 5µl 2x TaqMan™ Gene Expression Master Mix (Applied Biosystems, cat number 4369016) and 1.5µl distilled H_2O.

The PCR was performed and analysed on a 7900HT real time PCR system (Applied Biosystems). The DNA was quantified against a standard curve and normalized as a percentage of the input.

2.3.6. - Library Preparation for ChIP-Sequencing

To check the fragment size of the ChIP samples 2µl of sample was run on a HS D1000 ScreenTape on the Agilent 2200 Tapestation. As the fragment size of the samples was quite large a Covaris E-220 ultrasonicator was used to shear the DNA, as per manufacturer's instructions, to fragments of about 500bp in size. Samples were sonicated with a duty cycle of 5%, a Peak Incident Power of 105 with 200 cycles per burst for 65 seconds. This was repeated twice for each sample. Samples were ethanol precipitated. 12ul of 5M sodium chloride, 360ul of 100% ethanol and 1ul of glycogen were added to each tube. The solution was then incubated at -20°C for 30 minutes

before being centrifuged at 20,000g at 4°C for 10 minutes. The supernatant was removed and the pellet was washed in 70% ethanol. The tube was centrifuged at 20,000g at 4°C for 5 minutes, and the ethanol wash was then removed. The pellet was left to dry before the DNA was resuspended in 16µl ddH_20. The concentration was measured on the Qubit®, before libraries were prepared.

Libraries were prepared using a DNA SMART™ ChIP-Seq Kit for Illumina (Clontech, cat number 634865), according to the manufacturers instructions. The samples were heated to 94°C for 2 minutes to ensure that the DNA was single-stranded. The samples were incubated on ice for at least 2 minutes, and then 3.25µl of DNA SMART buffer and 0.75µl were added to each sample. The tubes were vortexed briefly to mix. The samples were then incubated on a thermal cycler under the following conditions:

37°C for 10 min
65°C for 5 min
4°C hold

1µl of DNA SMART T-Tailing Mix and 1µl of Terminal Deoxynucleotidyl Transferase were added to each tube, and vortexed. The samples were then incubated on a thermal cycler under the following conditions:

37°C for 20 min
70°C for 10 min
4°C hold

2µl of DNA SMART Poly(dA) Primer was added to each sample, and heated to 94°C for 1 minute before being incubated on ice for at least 2 minutes. 6µl of DNA SMART Buffer, 6µl of DNA SMART Oligonucleotide Mix and 4µl of SMARTScribe Reverse Transcriptase were added to each sample, and vortexed to mix. The samples were then incubated on a thermal cycler under the following conditions:

42°C for 90 min

70°C for 15 min

4°C hold

50µl of SeqAmp PCR Buffer (2x), 2µl of Forward PCR Primer (12.5M), 2µl of Reverse PCR Primer (12.5M) and 2µl of SeqAmp DNA Polymerase were added to each sample, and vortexed briefly to mix. Different reverse primers were used for each sample so that the samples can be differentiated after sequencing. The samples were then incubated on a thermal cycler under the following conditions:

94°C for 1 min

98°C for 15 sec
55°C for 15 sec
68°C for 30 sec
} 15 Cycles

4°C hold

The libraries were sized selected to enrich for sequences between 250 and 500bp. 75µl of AMPure XP beads (Beckman Coulter, cat number A62880) were added to each library, and mixed by pipetting at least 10 times. Samples were incubated for 8 minutes

at room temperature. The samples were placed on a Magnetic Separation Device for 10-20 minutes (or until the solution has cleared). 25µl AMPure XP beads were added to new PCR tubes. The sample tubes were left on the magnetic stand and the supernatant was transferred to the new PCR tubes containing the beads, and mixed by pipetting at least 10 times. Samples were incubated for 8 minutes at room temperature. The samples were placed on a Magnetic Separation Device for 10-20 minutes (or until the solution has cleared). The sample tubes were left on the magnetic stand and the supernatant was removed. 200µl of freshly made 80% ethanol was added to each sample, without disturbing the beads. After 30 seconds the supernatant was carefully removed. This ethanol wash was repeated once. The samples were incubated for 3-5 minutes (until the pellet was dry) at room temperature. Once the pellet was dry the tubes were removed from the magnetic stand, and 20µl of Library Elution Buffer was added to each tube. The samples were mixed by pipetting the beads up and down, and incubated for 5 minutes at room temperature. The samples were placed on a Magnetic Separation Device for 2 minutes or until the solution had cleared. The clear supernatant was transferred to a new tube and stored at -20°C.

2.3.7. - Quantification of ChIP Library

The libraries were quantified using qPCR comparing them against a set of standards using the KAPA SYBR® FAST ABI Prism qPCR kit (Anachem Ltd, cat number KK4835).

2.3.8. - ChIP-Seq Data Analysis

This was performed by Dr Nikolas Barkas.

Raw read quality was assessed with FASTQC (online reference available only:

http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Five base pairs were trimmed from the beginning of each read and 20 from the end as low sequence quality was observed in the corresponding sequencing cycles.

Sequence reads were aligned to the mouse reference genome (mm10) using Bowtie2 (v. 2.2.6 [148] [149]). Duplicates were identified with the Picard Toolkit (v 1.140) MarkDuplicates command and subsequently filtered. MACS2 (http://liulab.dfci.harvard.edu/MACS/ reference macs1) was used to identify peaks at the $q < 0.01$ significance level. Peaks that were identified in both replicates were used for subsequent analysis.

2.3.9. – Parental Allele-Specific Binding Analysis

This was performed by Dr Nikolas Barkas.

Parental allele-specific binding was assessed by binomial testing, using a custom bioinformatics pipeline. For performance reasons, only reads of interest, which overlapped the previously identified CTCF or RAD21 binding sites, were extracted from the SAM files and used for subsequent analysis.

Individual reads were assigned to one of the parental alleles using a custom Perl script, using the SAMtools Perl library. Each read was mapped as either derived from the reference sequence (Bl6) or from the FIX ME (MOLF_EiJ) allele on the basis of a SNP between the parental strains. If more than one SNP was present, the SNP with the best quality of read sequence was used. Reads were only considered for subsequent analysis if the Phred-scaled alignment mapping quality exceeded 20 and the base call quality at the SNP used for mapping of the read exceeded 5.

Paired reads were mapped to parental strains separately. As paired reads are not independent data points, when they were in disagreement (<1%) the read pair was assigned on the basis of the best SNP in either of the two reads.

Assigned reads were converted to maternally or paternally derived. Counts of maternal and paternal reads were obtained on a per-region basis using MySQL. Binding regions were only tested for parent-of-origin-specific expression if three or more reads could be mapped.

Parental allele-specific binding was assessed using a two-sided binomial test (implemented in R) of the maternal-versus-paternal allelic read counts. Regions were sorted by *P*-value score using MySQL. The genome-wide significance of *P*-values was assessed by means of Bonferroni correction.

2.4. – Chapter 4

2.4.1. - DNA Extraction from ES Cells

We received frozen cell pellets from our collaborators. The pellets were resuspended in 100µl of TE buffer (pH8.0). 1ml of Extraction Buffer was added to each sample, before they were incubated at 37°C for 1hr. 7µl of Proteinase K (to a final concentration of 100µg/ml) was added to each sample and mixed gently. The samples were then incubated at 50°C for 3hours, and swirled periodically. They were then cooled to room temperature. Phase lock tubes (VWR, cat number 713-2533) were prepared by centrifuging them at 13,000g at 4°C for 1 minute. One volume of phenol:chloroform

was added to the DNA in a phase lock tube and inverted to mix the solution. The solution was then centrifuged at 13,000RPM at 4°C for 10 minutes. The top aqueous layer was removed and transferred to a fresh microfuge tube. 0.25 volumes of 7.5M ammonium acetate and 2 volumes of 100% ethanol were added to each tube. The samples were centrifuged at 5,000g for 5 minutes. The supernatant was removed and the pellet was washed in 70% ethanol. The tube was centrifuged at 5,000g for 5 minutes, and the ethanol wash was then removed. The pellet was left to dry before the DNA was resuspended in 200µl TE buffer pH8.0. Samples were then placed on a rocking platform at room temperature overnight. Extracted DNA was stored at 4°C.

2.4.2. - Restriction Digest

9µg of DNA was digested with 3µl of enzyme and 12µl of appropriate buffer in a total volume of 114µl, at 37°C for 2hours. An additional 6µl of enzyme was added to each tube before incubating at 37°C overnight. The following morning the samples were heated to 65°C for 20 minutes, for all enzymes except *Spe*I which required heating to 80°C for 20 minutes, to inactivate the enzyme.

Digest for	Enzyme	NEBuffer
Fam13c probe	*Hind*III	2
Fam13c neo probe	*Stu*I	4
H13c probe	*Xba*I	4 + BSA (100µg/ml)
H13c neo probe	*Spe*I	4 + BSA (100µg/ml)

Table 2.2 – A summary of the restriction enzymes, and the conditions required for digestion. All restriction enzymes, BSA and buffers used were purchased from New England Biolabs®, see appendix 7.3. for details.

The DNA was concentrated into a suitable volume for loading on a gel. 300µl of 100% ethanol and 12µl of 3M sodium acetate was added to each sample. Samples were then incubated at 4°C for 30 minutes. They were then centrifuged at 13,000RPM at 4°C for 20 minutes. The supernatant was removed and the pellet was washed in 70% ethanol. The samples were centrifuged at 13,000RPM at 4°C for 10 minutes, and the ethanol wash was then removed. The pellet was left to dry before the DNA was resuspended in 25µl ddH_2O.

2.4.3. - Southern Blotting

Samples were loaded onto a 1% agarose gel and electrophoresed at 110V until the DNA migrated from the wells into the gel. Once this was achieved the voltage was lowered to 25V and the gel left to run overnight at room temperature. The gel was imaged with a ruler aligned to the ladder, before being washed in denaturation buffer for 30 minutes, shaking gently at room temperature. This was followed by two 15 minute washes in neutralisation buffer, again shaking gently at room temperature. A capillary blot was then set up to transfer the DNA in the agarose gel onto a Hybond-N+ membrane (Amersham, cat number RPN 203B), using 20x SSC. Once assembled this was left overnight (2-16hours). The location of the wells were marked on the membrane with a pencil. The membrane was then washed briefly in 2X SSC, before being left to dry for 2hours at 80°C.

The membrane was incubated in Church buffer containing 150µl of sheared Salmon sperm DNA rotating at 65°C overnight. The Church buffer was replaced with fresh Church buffer plus Salmon sperm DNA. 25ng of the DNA probe was prepared, by boiling for 5 minutes before being incubated on ice. 4µl High –prime mix (Sigma

Aldrich, cat number 11585592001) and 3µl α^{32} P – dCTP (at 10µCi/µl) were added to the probe. The labelled probe was incubated at 37°C for 1hr. 2µl of 0.2M EDTA was added, and the probe was incubated at 65°C for 10 minutes. The probe was transferred to a spin column and centrifuged at 1,100g for 4 minutes, before being incubated at 95°C for 10 minutes. The probe was added to the membrane, and incubated at 65°C overnight.

The membrane was washed twice in wash solution 1 at 65°C for 15 minutes. The membrane was monitored for radioactivity. If the signal was low and specific the membrane was then wrapped in saran wrap and placed in a cassette and left to develop. If the label was still non specific the membrane was then washed twice in wash solution 2 at 65°C for 1hour, checking the specificity of the label between washes.

2.4.4. - Long-Range PCR

An Expand Long Template PCR System kit was used (Roche, cat number 11681842001), according to the manufacturers instructions. 1µl DNA from the H13 knock-out Embryonic Stem cell samples was added to 1.5µl 10µM Forward primer, 1.5µl 10µM primer Reverse primer, 10mM dNTP, 5µl supplied buffer 1, 0.5µl DNA polymerase mix and 40µl H_2O. The reactions were then incubated on a PCR machine under the following conditions:

94°C for 2 min

94°C for 10 sec
X°C for 30 sec } 10 Cycles
68°C for 4 min

94°C for 15 sec

X°C for 30 sec } 20 Cycles

68°C for 4 min + 20sec/cycle

68°C for 20min

The annealing temperatures and buffers needed for each primer set are summarised in table 2.3.

Primers	Target	Annealing Temperature (°C)	Buffer	Template	Product Size (Kb)
H13 Screen F1 H13 Screen neo R1	5' end of construct (including *Mcts2*)	60	1	Knock-out ES cells	4.9
Conseq F10 H13 Screen R1	3' end of construct (including the neomycin resistance gene)	60	1	Knock-out ES cells	4.1

Table 2.3 – The primer name, region amplified, template, buffer and annealing temperature for each long range PCR reaction.

2.4.5. – Gel Separation of PCR Products and DNA Extraction

The products of the PCR reactions were electrophoresed on a 1% agarose gel at 100V for about 90minutes. The gels were then imaged and the appropriate bands extracted. The DNA was recovered using a MinElute Gel Extraction Kit (Qiagen™, cat number 28604) and eluted in 10µl EB buffer, according to the manufacturer's instructions. The DNA was quantified using a NanoDrop™.

2.4.6. - Ligation

The ligation reaction was set up on ice. Insert DNA was added to vector DNA, 1µl of 10x Rapid ligation buffer, 1µl of T4 DNA ligase (Promega™, cat number M1801) and

ddH_2O to a total of 10µl. The following equation was used to work out the concentration of insert to add. For most of the ligation reactions 50ng of vector DNA was used. The ligation was incubated at 4°C overnight.

$$\frac{\text{ng of vector} \times \text{kb size of insert}}{\text{kb size of vector}} \times \text{insert:vector molar ratio} = \text{ng of insert to add}$$

2.4.7. – Transformation into Chemically Competent Cells

3µl of the ligation reaction was then added to 50µl competent *E.coli* cells on ice, and swirled to mix. The cells were incubated on ice for 30 minutes before being heat shocked at 42°C for 45 seconds. The cells were then returned to the ice for 2 minutes. 100µl of S.O.C. medium (Invitrogen, cat number 46-0821) was aseptically added to the cells, and they were incubated at 37°C for 1hour shaking. After 1hour the cells were plated out onto Ampicillin (200µl of 50µg/ml in 100mls), IPTG (8.4µl of 1M in 100mls), X-Gal (100µl of 40mg/ml in 100mls) plates and left overnight at 37°C. Colonies were then picked and incubated in 5ml LB and Ampicillin (0.05µg/ml) overnight at 37°C shaking.

2.4.8. - DNA Extraction from Bacterial Cultures

DNA was extracted from a single colony grown overnight in 5mls LB and appropriate antibiotic, using a Promega SV Miniprep Wizard Plus Kit (Promega, cat number A1460) according to the manufacturers protocol.

2.4.9. - Sequencing

To prepare the samples for sequencing 1µl of DNA was added to 2µl of 5x sequencing buffer, 0.4µl 10µM primer, 6.1µl H_2O and 0.5µl Big Dye Terminator v3.1 (Invitrogen, cat number 4337454). This was added to each well of a 96 well plate. The plate was then run on a PCR machine under the following conditions:

96°C for 1 min

96°C for 30 sec
58°C for 15 sec } 30 Cycles
62°C for 1 min

To precipitate the DNA 30µl of 100% ethanol and 1µl of 3M sodium acetate was added to each well. The plate was incubated at 4°C for 20 minutes. The plate was centrifuged at 3060g at 4°C for 20 minutes, before the ethanol was removed and replaced with 70% ethanol. The plate was incubated at 4°C again, this time for 5 minutes, before being centrifuged at 3060g at 4°C for 10 minutes. The ethanol was removed and the plate was left to dry at room temperature for 20 minutes. The pellets were then resuspended in 10µl Hi-Di formamide (Applied Biosystems, cat number 4311320), and 10µl 1mM EDTA was added to empty wells. The plate was incubated at 94°C for 2 minutes and then cooled on ice before being sequenced using a 3730xl Sanger sequencer machine. Sequence traces were analysed using Sequencer.

2.4.10. – Generation of Constructs

There are three constructs to assess transcriptional interference at the *H13/Mcts2* locus. All of the constructs contain UCOE (ubiquitous chromatin opening element) allowing

transcription from the construct regardless of its integration site in the genome. The three constructs each contain different combinations of the introns and exons of *H13* upstream of *Mcts2*. In place of *Mcts2* the three contain a minimal promoter linked to the expression of mCherry. This promoter will be under the control of a tetracycline inducible element. All three constructs contain the sequence of intron 4 of *H13* immediately downstream of *Mcts2* and the complete sequence of exon 5, allowing splicing to occur from exon 4 onto exon 5. All of the constructs also contain eGFP located after exon 5, as a readout of transcription through exon 5. When the construct is transfected into cells it will allow us to determine if it is transcription of the internal gene that stimulates the use of upstream poly (A) sites, through the readout of either mCherry or eGFP. I would expect to see eGFP expression when the cells aren't exposed to tetracycline, as there should be no transcription from the minimal promoter, allowing use of the down-stream poly (A) sites. In the presence of tetracycline when the minimal promoter is stimulated I would expect to see expression of mCherry and a decrease in the expression of GFP as some (if not all) transcripts terminate at a poly (A) site upstream of the minimal promoter.

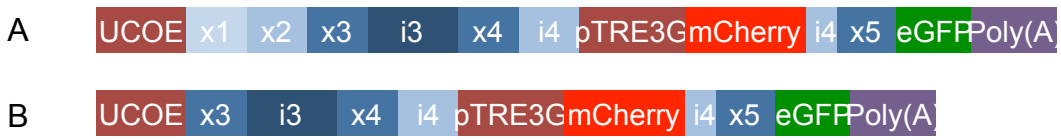


Figure 2.1 – Construct for investigating the affect of an internal promoter on the transcription of the host gene. The star shows the location of the poly (A) site used to generate the shorter transcripts in the *H13/ Mcts2* locus that this construct is based on. x1, x2, x3, x4 and x5 are the corresponding exons from *H13*. i1, i2,i3 and i4 are based on the corresponding introns in *H13*. Due to the large size of i1 and i2 I have used the first and last 250bp of each of these introns.

The constructs were generated through a series of cloning steps (summarised in Figure 4.8 and section 2.4.10.1 later in the Materials and Methods) requiring a large number of unique restriction sites allowing us to control the addition of each fragment to the construct, and if necessary, allow its easy removal, without disrupting the rest of the construct. Cloning steps two and three (Figures 4.9 and 4.10 respectively) required the replacement of the multiple cloning site (MCS) of two vectors with 'linker' sequences designed to contain the required restriction sites. As these steps required the insertion of a small (85-150bp) fragment into the vector a large amount of DNA was digested, and the digests were electrophoresed through low melting point (LMP) agarose gels to maximise the yield of DNA recovered. These steps were performed in parallel because different vectors were being used for these steps.

Steps six (Figure 4.11), eight (Figure 4.12) and nine (Figure 4.13), and assembling the *H13* sequence (Figure 4.13) required PCR amplification of genomic or cDNA with primers containing the required restriction enzyme sites for inserting the products into the correct place in the vectors.

Steps four (Figure 4.12), 12 (Figure 4.14) and 14 (Figure 4.15) required the use of only one restriction enzyme for the insertion of the fragments into the vector. This meant that in addition to checking that the fragment had been incorporated into the vector, it was also necessary to check that the insert was correctly orientated. To screen for this we used a double restriction enzyme digest, with one restriction site located towards one end of the insert and another located in the vector. The different sizes generated by the digest allowed us to determine which clones contained the insert in the correct orientation.

Once all steps were complete the entire construct was sequenced enabling us to verify that we had successfully generated both versions of the construct as designed.

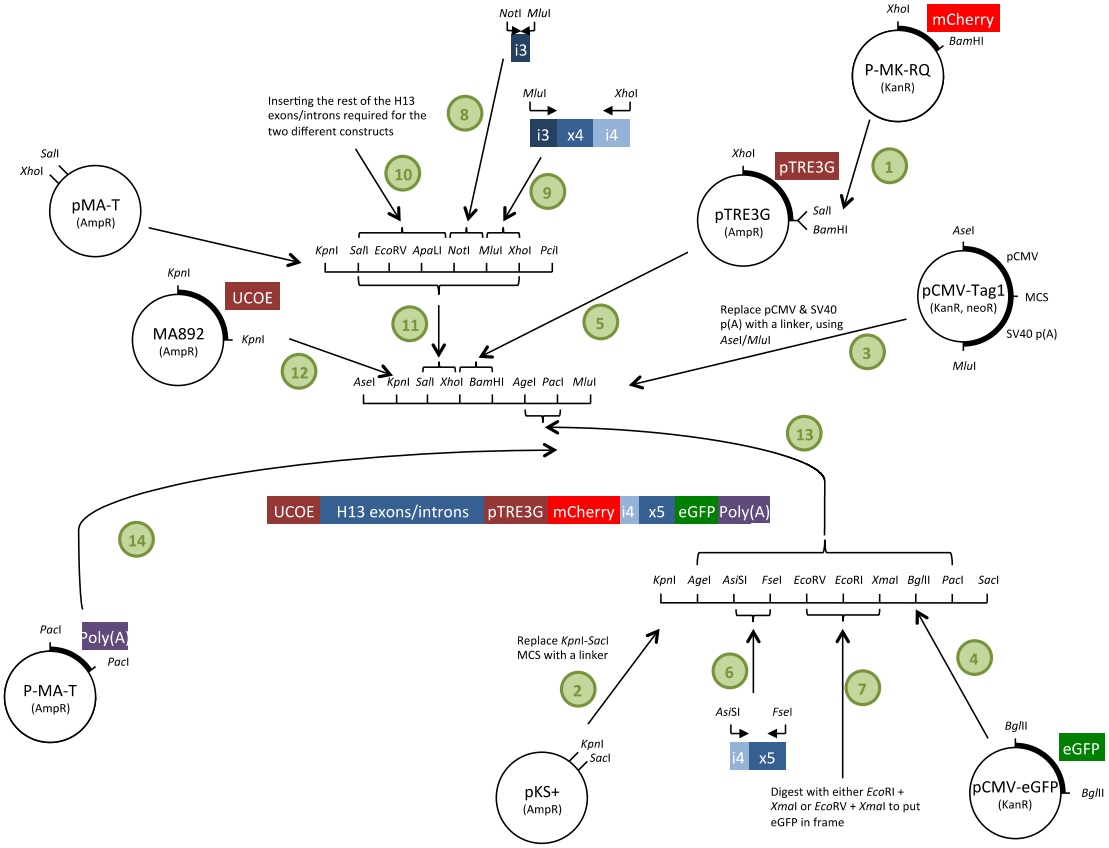


Figure 2.2 – Summary of the cloning steps required to generate the constructs. The restriction enzymes required for each cloning step are shown. The antibiotic resistance of each plasmid has been included. UCOE is an Ubiquitously acting Chromatin Opening Element. mCherry is a red fluorescent protein. eGFP is an enhanced Green Fluorescent Protein. Green circles: step numbers, representing the order in which cloning steps were performed.

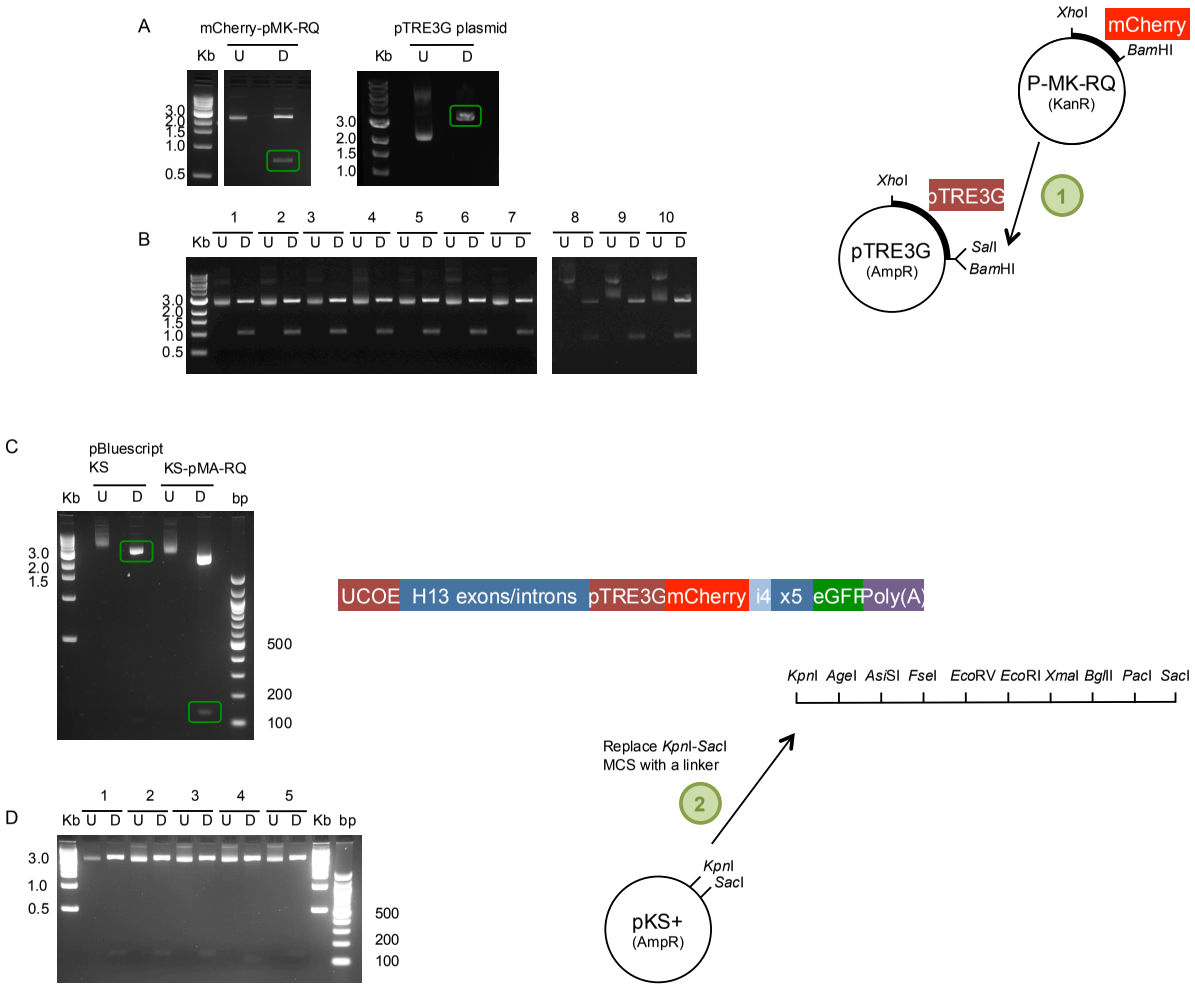


Figure 2.3 - Cloning steps 1 and 2 to generate constructs for the study of transcriptional interference. A – mCherry-pMK-RQ was digested with *Bam*HI-HF and *Xho*I, and the pTRE3G containing plasmid (Clontech) was digested with *Bam*HI-HF and *Sal*I. The fragments were electrophoresed through agarose, and DNA from the bands of interest, highlighted in green, were extracted for ligation. DNA from the band corresponding to the 746bp mCherry fragment and the band corresponding to the linearised pTRE3G plasmid were extracted. B – After transformation into competent cells 10 clones were picked, DNA was extracted, and digested with *Bam*HI-HF and *Xho*I to verify successful uptake of the mCherry sequence. C – pBluescript-KS and KS-pMA-RQ were both digested with *Kpn*I and *Sac*I. The samples were electrophoresed through agarose, and DNA from the bands of interest, highlighted in green, were extracted for ligation. DNA from the band corresponding to the 148bp KS linker fragment and the band corresponding to the linearised pBluescript-KS were extracted. D – After cloning 5 clones were digested with *Kpn*I and *Sac*I to verify successful uptake of the KS linker sequence. U = Undigested sample, D = Digested sample.

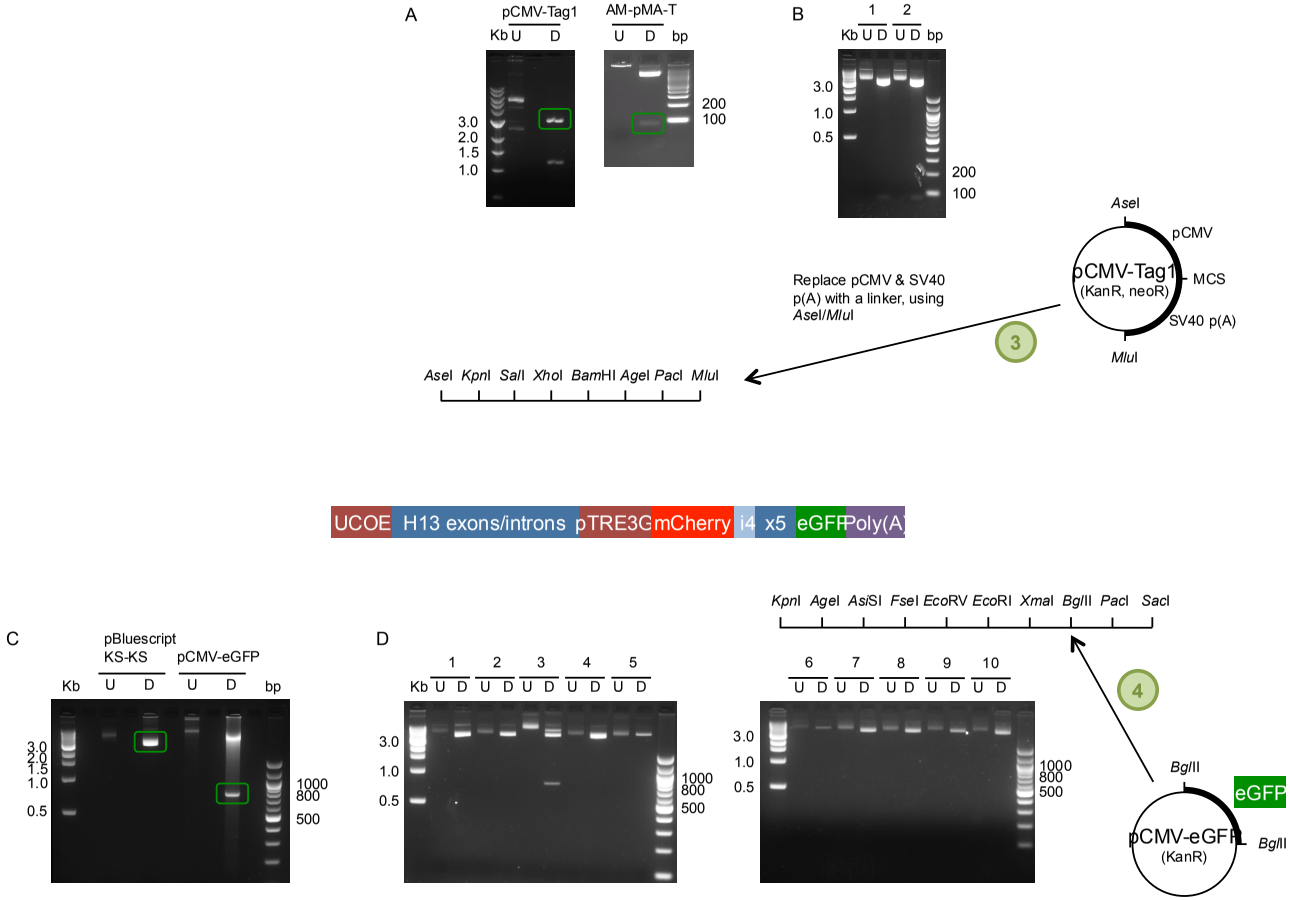


Figure 2.4 - Cloning steps 3 and 4 to generate constructs for the study of transcriptional interference. A – pCMV-Tag1 and AM-pMA-T were digested with *Ase*I and *Mlu*I. The fragments were electrophoresed through agarose, and DNA from the bands of interest, highlighted in green, were extracted for ligation. DNA from the band corresponding to the 84bp AM linker fragment and the band corresponding to the linearised pCMV-Tag1 plasmid were extracted. B – After transformation into competent cells 2 clones were picked, DNA was extracted, and digested with *Ase*I and *Mlu*I to verify successful uptake of the AM linker sequence. C – pBluescript KS-KS and pCMV-eGFP were both digested with *Bgl*II. The samples were electrophoresed through agarose, and DNA from the bands of interest, highlighted in green, were extracted for ligation. DNA from the band corresponding to the 800bp eGFP fragment and the band corresponding to the linearised pBluescript KS-KS were extracted. D – After cloning 10 clones were digested with *Bgl*II to verify successful uptake of the eGFP fragment. U = Undigested sample, D = Digested sample.

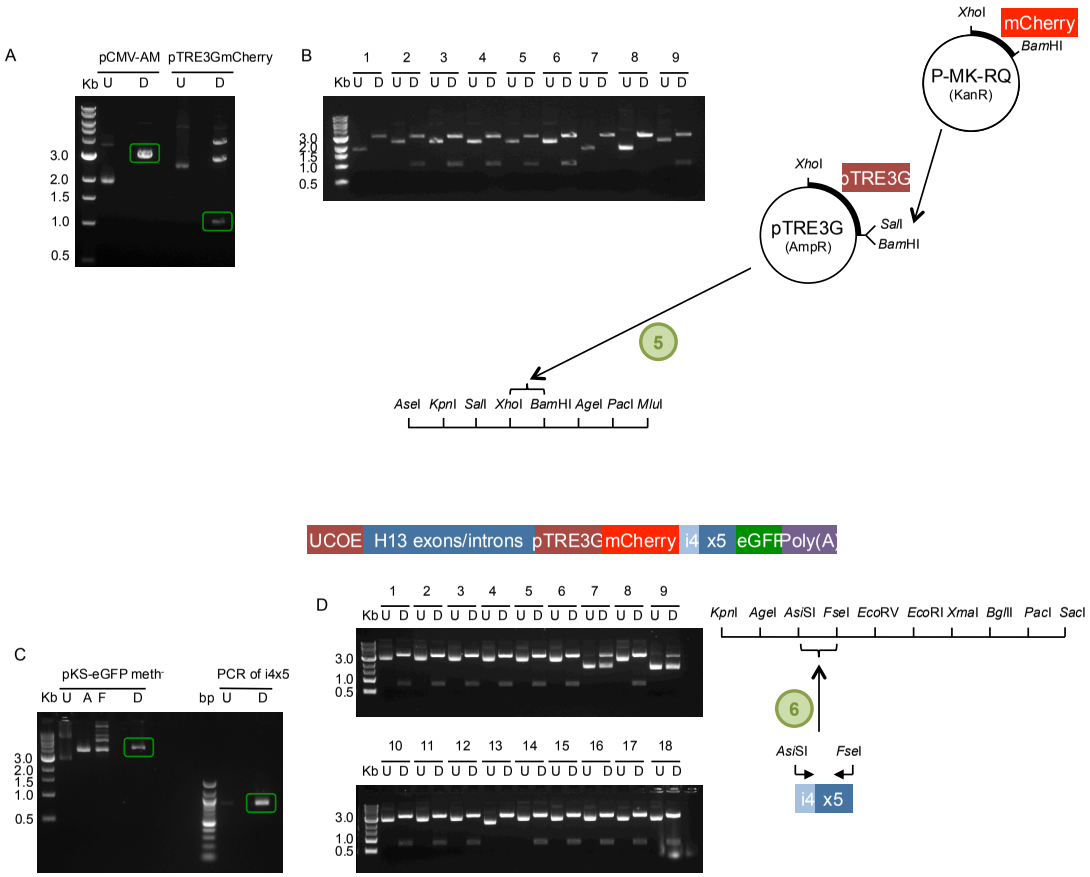


Figure 2.5 - Cloning steps 5 and 6 to generate constructs for the study of transcriptional interference. A – pCMV-AM and pTRE3GmCherry were digested with *BamH*I-HF and *Xho*I. The fragments were electrophoresed through agarose, and DNA from the bands of interest, highlighted in green, were extracted for ligation. DNA from the band corresponding to the 1.126Kb pTRE3GmCherry fragment and the band corresponding to the linearised pCMV-AM plasmid were extracted. B – After transformation into competent cells 9 clones were picked, DNA was extracted, and digested with *BamH*I-HF and *Xho*I to verify successful uptake of the pTRESGmCherry sequence. C – pKS-eGFP meth$^-$ and H13 intron 4 exon 5 (generated by PCR) were both digested with *Asi*SI and *Fse*I. The samples were electrophoresed through agarose, and DNA from the bands of interest, highlighted in green, were extracted for ligation. DNA from the band corresponding to the 800bp H13 intron 4 exon 5 fragment and the band corresponding to the linearised pKS-eGFP meth$^-$ were extracted. D – After cloning 18 clones were digested with *Asi*SI and *Fse*I to verify successful uptake of the H13 intron 4 exon 5 fragment. U = Undigested sample, D = Digested sample, A = Digested with *Asi*SI, F = Digested with *Fse*I.

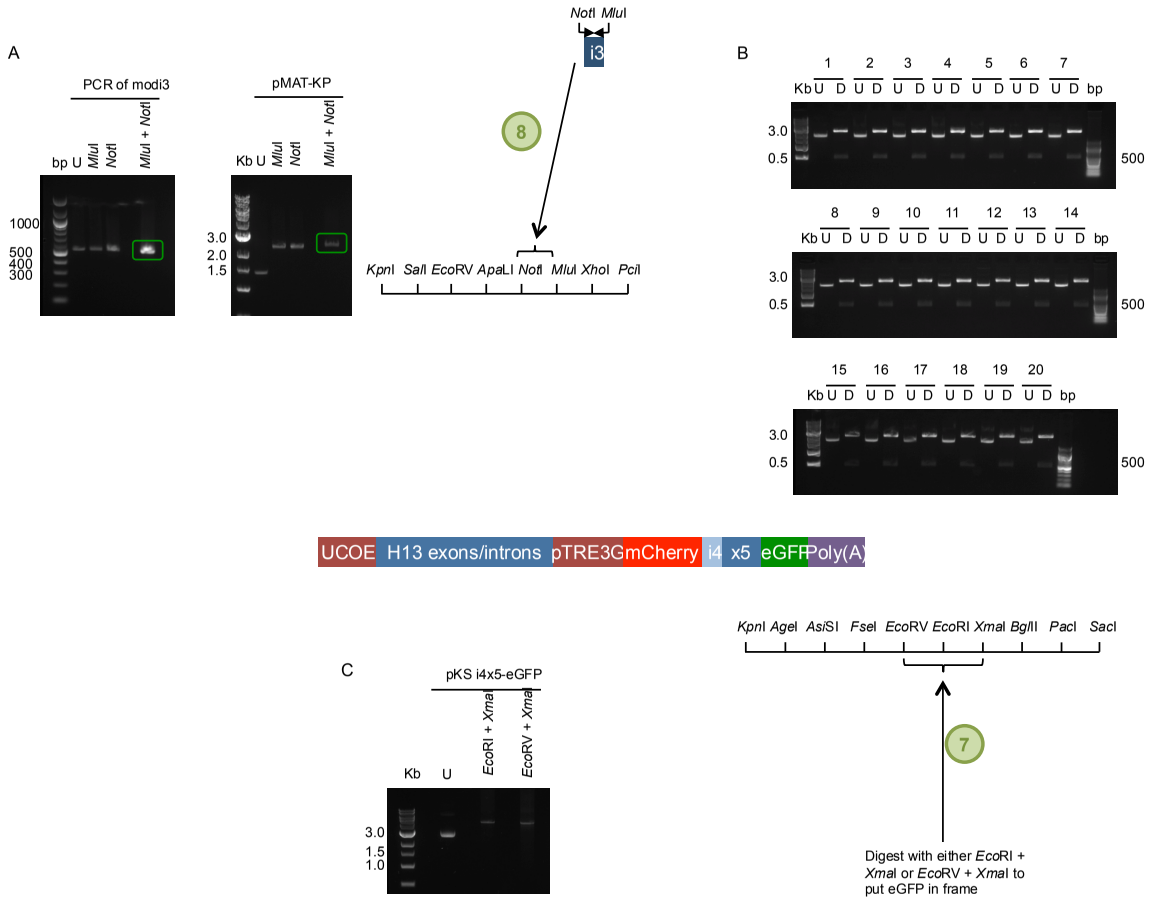


Figure 2.6 - Cloning steps 7 and 8 to generate constructs for the study of transcriptional interference. A – pMAT-KP and H13 modified intron 3 (generated by PCR) were digested with *Not*I and *Mlu*I. The fragments were electrophoresed through agarose, and DNA from the bands of interest, highlighted in green, were extracted for ligation. DNA from the band corresponding to the 550bp H13 modified intron 3 fragment and the band corresponding to the linearised pMAT-KP plasmid were extracted. B – After transformation into competent cells 20 clones were picked, DNA was extracted, and digested with *Not*I and *Mlu*I to verify successful uptake of the modified H13 intron 3 sequence. C – pKS i4x5-eGFP was digested with either *EcoR*I and *Xma*I or *EcoR*V and *Xma*I to remove less than 10bp to put eGFP into frame in both the final constructs. The samples were electrophoresed through agarose, and DNA from the bands of interest, highlighted in green, were extracted for ligation. DNA from the bands corresponding to the linearised pKS i4x5-eGFP were extracted. After transformation into competent cells 20 clones were picked, DNA was extracted and sequenced to check for the removal of the bases required to put eGFP into frame in the final constructs. U = Undigested sample, D = Digested sample.

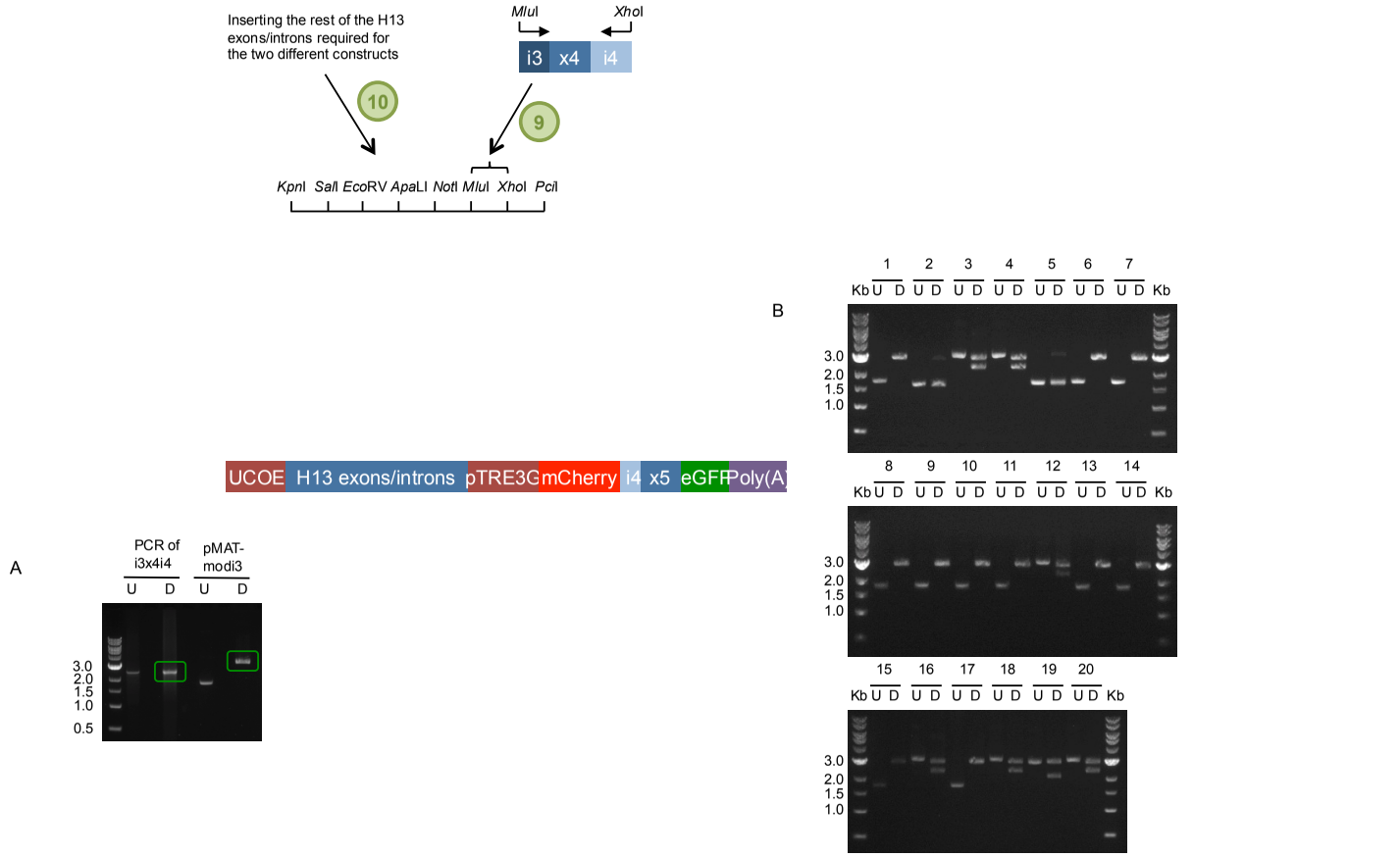


Figure 2.7 - Cloning steps 9 and 10 to generate constructs for the study of transcriptional interference. A – pMAT-modi3 and H13 intron 3 exon 4 intron 4 (generated through PCR) were digested with *Mlu*I and *Xho*I. The fragments were electrophoresed through agarose, and DNA from the bands of interest, highlighted in green, were extracted for ligation. DNA from the band corresponding to the 2.4Kb H13 intron 3 exon 4 intron 4 fragment and the band corresponding to the linearised pMAT-modi3 plasmid were extracted. B – After transformation into competent cells 20 clones were picked, DNA was extracted, and digested with *Mlu*I and *Xho*I to verify successful uptake of the H13 intron 3 exon 4 intron 4 sequence. U = Undigested sample, D = Digested sample.

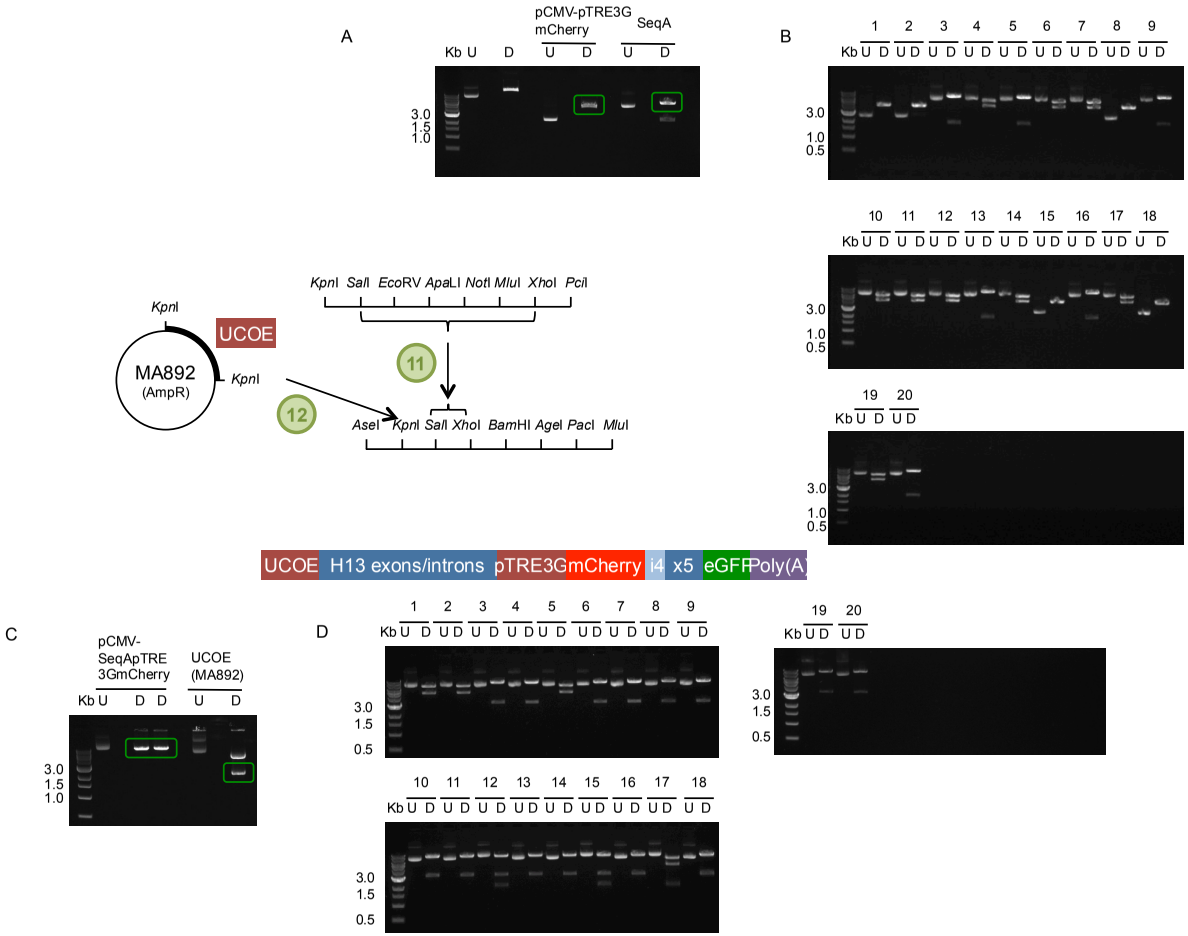


Figure 2.8 - Cloning steps 11 and 12 to generate constructs for the study of transcriptional interference. A – pCMV-pTRE3GmCherry and the sequence of H13 for construct A were digested with *Sal*I-HF and *Xho*I. The fragments were electrophoresed through agarose, and DNA from the bands of interest, highlighted in green, were extracted for ligation. DNA from the band corresponding to the 6Kb sequence of H13 for construct A fragment and the band corresponding to the linearised pCMV-pTRE3GmCherry plasmid were extracted. B – After transformation into competent cells 20 clones were picked, DNA was extracted, and digested with *Pci*I to verify successful uptake of the sequence of H13 for construct A. C – UCOE (MA892) and pCMV-SeqApTRE3GmCherry were both digested with *Kpn*I. The samples were electrophoresed through agarose, and DNA from the bands of interest, highlighted in green, were extracted for ligation. DNA from the band corresponding to the 2.2Kb UCOE fragment and the band corresponding to the linearised pCMV-SeqApTRE3GmCherry were extracted. D – After cloning 20 clones were digested with *Pac*I and *Xba*I to verify successful uptake of the UCOE fragment. U = Undigested sample, D = Digested sample.

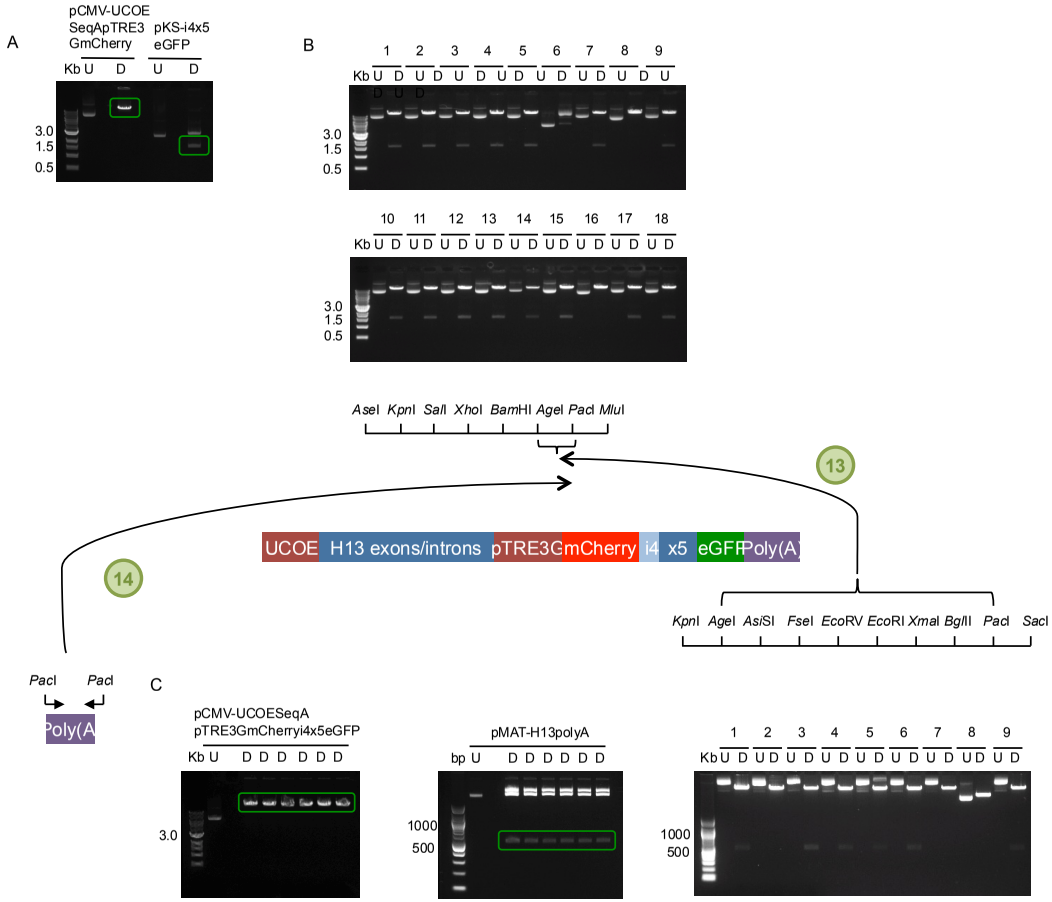


Figure 2.9 - Cloning steps 13 and 14 to generate constructs for the study of transcriptional interference. A – pCMV-AM and pKS-i4x5eGFP were digested with *Age*I and *Pac*I. The fragments were electrophoresed through agarose, and DNA from the bands of interest, highlighted in green, were extracted for ligation. DNA from the band corresponding to the 1.6Kb H13 intron 4 exon 5 eGFP fragment and the band corresponding to the linearised pCMV-UCOESeqApTRE3GmCherry plasmid were extracted. B – After transformation into competent cells 18 clones were picked, DNA was extracted, and digested with *Age*I and *Pac*I to verify successful uptake of the H13 intron 4 exon 5 eGFP sequence. C – pCMV-UCOESeqApTRE3GmCherryi4x5eGFP pMAT-H13polyA were both digested with *Pac*I. The samples were electrophoresed through agarose, and DNA from the bands of interest, highlighted in green, were extracted for ligation. DNA from the band corresponding to the 600bp H13polyA fragment and the band corresponding to the linearised pCMV-UCOESeqApTRE3GmCherryi4x5eGFP were extracted. D – After cloning 9 clones were digested with *Pac*I to verify successful uptake of the H13polyA fragment. U = Undigested sample, D = Digested sample.

2.4.10.1. – Source of Components

mCherry in pMK-RQ, H13i3 modified sequence in pMA-T, AM linker sequence in pMA-T, KP linker sequence in pMA-T, BH linker sequence in pMK-RQ, KS linker sequence in pMA-RQ, H13polyA sequence in pMA-T and the x1i3 sequence in pMK-RQ were generated by GeneArt AG (Life Technologies^{TM}). The pTRE3G promoter sequence was supplied in the Tet-On 3G Inducible Expression System (Clontech Laboratories Inc, cat number 631168) kit. An aliquot of pKS+ was generously donated by Dr Kirupa Sathasivam, a non-clinical research fellow in the Department of Medical and Molecular Genetics. pCMV-eGFP and pCMV-Tag1 were generated by Dr Mike Cowley, a post-doc in the lab. All the intronic and exonic sequences in the construct (labelled respectively as i and x in Figure 2.2) came from BxC new born mice.

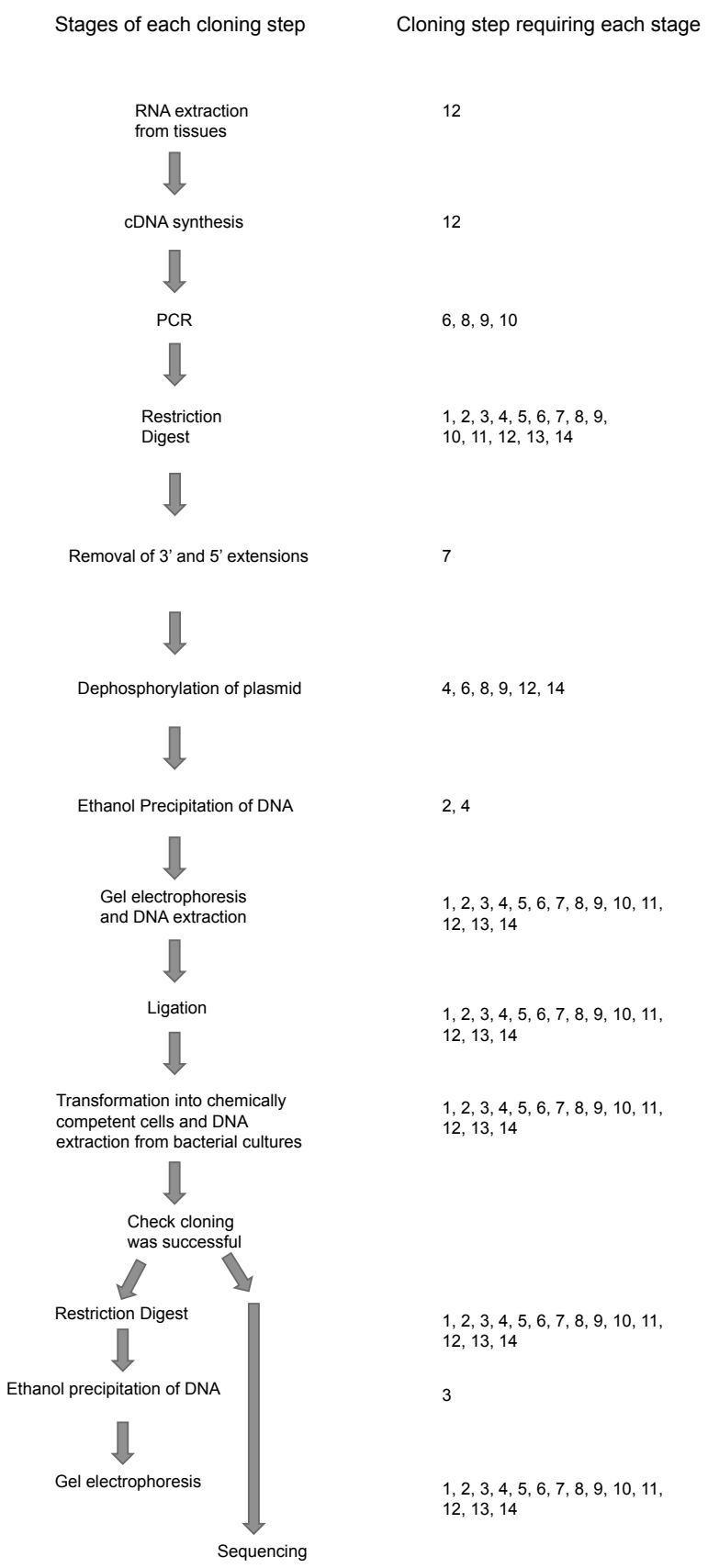


Figure 2.10 – Flow diagram summarising the stages required for each cloning step.

2.4.10.2. - RNA Extraction

An RNeasy Mini Kit (Qiagen, cat number 74104) was used according to manufacturers instructions. 600µl of buffer RLT was added to 30mg of tissue, and homogenised for 30 seconds using a rotor-stator homogeniser. The lysate was centrifuged at full speed for 3 minutes. The supernatant was transferred to a fresh microfuge tube. 1x the sample volume of 70% ethanol was added, and pipetted to mix the two well. Up to 700µl of the sample was transferred to an RNeasy spin column, placed in a 2ml collection tube. The sample was centrifuged at 10,000 RPM for 15 seconds, and the flow-through was discarded. This was repeated until the entire sample had been passed through the column. 350µl of buffer RW1 was added to the column, and the column was centrifuged at 10,000 RPM for 15 seconds, and the flow-through discarded. 10µl of DNase I stock solution was diluted in 70µl Buffer RDD, and mixed by inverting the tube. The DNase I incubation mix was added to the column membrane and incubated on the bench for 15 minutes. 350µl Buffer RW1 was added to the column, and the column was centrifuged at 10, 000 RPM for 15 seconds, and the flow-through discarded. 500µl Buffer RPE was added to the column, and the column was centrifuged at 10, 000 RPM for 15 seconds, and the flow-through discarded. 500µl Buffer RPE was added to the column, and the column was centrifuged at 10,000 RPM for 2 minutes. The column was moved to a fresh 2ml collection tube, and centrifuged at full speed for 1 minute. The column was moved to a fresh 1.5ml microfuge tube. 40µl RNase-free water was added directly to the column, and incubated on the bench for 1 minute. The sample was centrifuged at 10,000 RPM for 1 minute. The RNA was quantified using a NanoDrop™.

2.4.10.3. - cDNA Synthesis

1µg of RNA was added to 1µl 10mM dNTP's (Life Technologies™, cat number R0191), 1µl oligo dT primer (500µg/µl) (Life Technologies™, cat number 18418020), and nuclease-free water was added to a total volume of 13µl. The RNA was heated to 65°C for 5 minutes. The RNA was incubated on ice for at least 1 minute while 4µl of 5x first strand buffer, 1µl of 0.1M DTT and 1µl of nuclease-free water were added. The 5 x first strand buffer and 0.1M DTT were supplied with the Superscript II RT. The samples were then incubated at 42°C for 2 minutes. 1µl of Superscript II RT (Invitrogen, cat number 18064-014) was added, and the samples incubated at 42°C for 50 minutes. The samples were heated to 70°C for 15 minutes, and the cDNA generated was stored at -20°C.

2.4.10.4. – DNA Amplification

2.4.10.4.1. – PCR

1µl DNA was added to 0.5µl 10µM Forward primer, 0.5µl 10µM Reverse primer, and 18µl 1.1x Reddy Mix (Thermo Scientific, cat number AB-0608/LD). The reactions were then incubated on a PCR machine under the following conditions:

94°C for 2 min
94°C for 30 sec
X°C for 30 sec } X Cycles
72°C for 1 min
72°C for 10min

DNA was amplified under the conditions specified in Table 2.4.

Primers	Target	Annealing Temperature (°C)	Cycle number	Template	Product Size (bp)
H13i4/x5 F1 H13i4/x5 R	i4x5	55	35	gDNA B6 adult heart	825
SH_modi3_F SH_modi3_R1	Modified i3	54	35	Modified H13i3	550

Table 2.4 - The primer name, region amplified, template, annealing temperature and cycle number for each PCR reaction.

2.4.10.4.2. – Long-Range PCR

According to the protocol in section 2.4.4. with the following alterations.

The extension time for the PCR is 2 minutes, instead of 4 minutes. The cycle number for the second round of PCR is 25, instead of 20. The annealing temperatures and buffers needed for each primer set are summarised in table 2.5.

Primers	Target	Annealing Temperature (°C)	Buffer	Template	Product Size (Kb)
SH_i3-i4_F1 SH_i3-i4_R	i3-i4	56	1	BxC nb Brain	2.4
SH_i3_F1 SH_i3_R	i3	58	3	BxC nb Brain	1.6
SH_x3-i3_F SH_i3_R	x3-i3	58	1	BxC nb Brain	2.7

Table 2.5 – The primer name, region amplified, template, buffer and annealing temperature for each long range PCR reaction.

2.4.10.5. – DNA Extraction from PCR Products

DNA was extracted from PCR products using the MinElute PCR Purification Kit (Qiagen™, cat number 28004). 5 volumes of buffer PBI were added to 1 volume of the PCR reaction, and placed in a MinElute column in a 2ml collection tube. The column was centrifuged at 10,000g for 1 minute, and the flow-through discarded. 750µl of buffer PE was added to the column, and the column was centrifuged at 10,000g for 1

minute, and the flow-through discarded. The column was centrifuged at 10,000g for 1 minute. The column was moved to a fresh 1.5ml microfuge tube, and 10µl of buffer EB was added to the centre of the membrane of the column. The column was left to stand for 1 minute, and then centrifuged at 10,000g for 1 minute. The DNA was quantified using a NanoDrop™.

2.4.10.6. – Restriction Digest

DNA was digested for 1hour at 37°C, unless otherwise specified. Samples were digested in 1/10 of the total volume of appropriate NEBuffer (with BSA if required), using 1µl each of appropriate restriction enzymes, unless otherwise specified.

		Amount of DNA	Restriction Enzymes	NEBuffer	Incubation time
1 – Cloning mCherry into the pTRE3G plasmid	mCherry-pMK-RQ	500ng	*BamH*I-HF and *Xho*I 1µl of each	4 and BSA (100µg/ml)	1 hr
	pTRE3G	1µg	*BamH*I-HF and *Sal*I 1µl of each	4	1 hr
2 – Cloning the KS linker sequence into pBluescript KS	KS-pMA-RQ	3µg	*Kpn*I and *Sac*I 3µl of each	1 and BSA (100µg/ml)	Overnight
	pBluescript KS	2µg	*Kpn*I and *Sac*I 3µl of each	1 and BSA (100µg/ml)	Overnight
3 – Cloning the AM linker into pCMV-Tag1	AM-pMA-T	500ng	*Ase*I and *Mlu*I 1µl of each	3	1 hr
	pCMV-Tag1	1µg	*Ase*I and *Mlu*I 1µl of each	3	1 hr
4 – Cloning eGFP into pBluescript KS-KS (product of step 2)	pCMV - eGFP	2µg	*Bgl*II 6µl	3	1 hr
	pBluescriptKS-KS	2µg	*Bgl*II 6µl	3	1 hr
5 – Cloning pTRE3GmCherry (product of step 1) into pCMV-AM (product of step 3)	pTRE3GmCherry	1µg	*BamH*I-HF and *Xho*I 2µl of each	4 and BSA (100µg/ml)	1 hr
	pCMV-AM	1µg	*BamH*I-HF and *Xho*I 2µl of each	4 and BSA (100µg/ml)	1 hr
6 – Cloning i4x5 into pBluescript KS-eGFP (product of step 4)	i4x5	1µg	*AsiS*I and *Fse*I 1µl of each	4 and BSA (100µg/ml)	Overnight
	pBluescriptKS-eGFP	1µg	*AsiS*I and *Fse*I 1µl of each	4 and BSA (100µg/ml)	Overnight
7 – Putting eGFP into frame (using the product of step 6)	pBluescript i4x5-eGFP	1µg	*EcoR*I and *Xma*I 1µl of each	4 and BSA (100µg/ml)	Overnight
	pBluescript i4x5-eGFP	1µg	*EcoR*V and *Xma*I 1µl of each	4 and BSA (100µg/ml)	Overnight

8 – Cloning modified H13i3 into pMAT-KP	Modified H13i3	2µg	*Mlu*I and *Not*I 2µl of each	3.1	2 hrs
	pMAT-KP	1µg	*Mlu*I and *Not*I 2µl of each	3.1	2 hrs
9 – Cloning i3x4i4 into pMAT-modi3 (product of step 8)	i3i4	2µg	*Mlu*I and *Xho*I 2µl of each	3.1	1 hr
	pMAT-modi3	1µg	*Mlu*I and *Xho*I 2µl of each	3.1	1 hr
11 – Cloning H13 sequence into pCMV-modi3i3i4pTRE3GmCherry (product of step 10)	pMAT-H13 sequence	1µg	*Sal*I-HF and *Xho*I 1µl of each	3.1	1 hr
	pCMV-pTRE3GmCherry	1µg	*Sal*I-HF and *Xho*I 1µl of each	3.1	1 hr
12 – Cloning UCOE into pCMV-H13SeqpTRE3GmCherry (product of step 11)	MA892	2µg	*Kpn*I 2µl	1 and BSA (100µg/ml)	2 hrs
	pCMV-H13SeqpTRE3GmCherry	2µg	*Kpn*I 2µl	1 and BSA (100µg/ml)	2 hrs
13 – Cloning i4x5eGFP into pCMV-UCOEH13SeqpTRE3GmCherry (product of step 12)	pBluescript i4x5-eGFP	2µg	*Age*I and *Pac*I 2µl of each	1 and BSA (100µg/ml)	2 hrs
	pCMV-UCOEH13SeqpTRE3GmCherry	2µg	*Age*I and *Pac*I 2µl of each	1 and BSA (100µg/ml)	2 hrs
14 – Cloning poly(A) into pCMV-UCOEH13SeqpTRE3GmCherryi4x5eGFP (product of step 13)	pMAT-H13polyA	2µg	*Pac*I 2µl	CutSmart	2 hrs
	pCMV-UCOEH13SeqpTRE3GmCherryi4x5eGFP	2µg	*Pac*I 2µl	CutSmart	2 hrs

Table 2.6 – Summary of digest conditions for generating fragments for ligation. All restriction enzymes, BSA and buffers used were purchased from New England Biolabs®, see appendix 7.3. for details.

Cloning Step	Digest			
	Amount of DNA	Restriction Enzymes	NEBuffer	Incubation time
1 – Cloning mCherry into the pTRE3G plasmid	300ng	*BamH*I-HF and *Xho*I 1µl of each	4 and BSA (100µg/ml)	1 hr
2 – Cloning the KS linker sequence into pBluescript KS	300ng	*Kpn*I and *Sac*I 1µl of each	1 and BSA (100µg/ml)	Overnight
3 – Cloning the AM linker into pCMV-Tag1	3µg	*Ase*I and *Mlu*I 3µl of each	3	1 hr
4 – Cloning eGFP into pBluescript KS-KS (product of step 2)	1µg	*Bgl*II 6µl	3	1 hr
5 – Cloning pTRE3GmCherry (product of step 1) into pCMV-AM (product of step 3)	300ng	*BamH*I-HF and *Xho*I 1µl of each	4 and BSA (100µg/ml)	1 hr
6 – Cloning i4x5 into pBluescript KS-eGFP (product of step 4)	500ng	*AsiS*I and *Fse*I 1µl of each	4 and BSA (100µg/ml)	2 hrs
8 – Cloning modified H13i3 into pMAT-KP	500ng	*Not*I and *Mlu*I 1µl of each	3.1	2 hrs
9 – Cloning i3x4i4 into pMAT-modi3 (product of step 8)	500ng	*Mlu*I and *Xho*I 1µl of each	3.1	1 hr
11 – Cloning H13 sequence into pCMV-modi3i3i4pTRE3GmCherry (product of step 10)	500ng	*Pci*I 1µl	3.1	1 hr
12 – Cloning UCOE into pCMV-H13SeqpTRE3GmCherry (product of step 11)	500ng	*Pac*I and *Xba*I 1µl of each	CutSmart	1 hr
13 – Cloning i4x5eGFP into pCMV-UCOEH13SeqpTRE3GmCherry (product of step 12)	500ng	*Age*I and *Pac*I 1µl of each	1 and BSA (100µg/ml)	1 hr

14 – Cloning poly(A) into pCMV-UCOEH13SeqpTRE3GmCherryi4x5eGFP (product of step 13)	500ng	*Pac*I	CutSmart	1 hr

Table 2.7 – Summary of digests to check that the cloning was successful. All restriction enzymes, BSA and buffers used were purchased from New England Biolabs®, see appendix 7.3. for details.

2.4.10.7. – Removal of 3' and 5' Extensions

4µl of Mung Bean Nuclease buffer and 1µl of Mung Bean Nuclease (New England Biolabs®, cat number M0250) were added to the digested pBluescript i4x5-eGFP, to a total of 40µl. The sample's were incubated for 30 minutes at 30°C. To inactivate the nuclease the samples were subjected to a phenol:chloroform extraction.

Phase lock tubes (VWR, cat number 713-2533) were prepared by centrifuging them at 13,000g at 4°C for 1 minute. One volume of phenol:chloroform (40µl) was added to the DNA in a phase lock tube and inverted to mix the solution. The solution was then centrifuged at 13,000RPM at 4°C for 10 minutes. The top aqueous layer was removed and transferred to a fresh 1.5ml microfuge tube. The DNA was then extracted using an ethanol precipitation.

2.4.10.8. – Dephosphorylation of Vector

To stop the vector self-ligating, the digested sample was treated with Antarctic Phosphatase (New England Biolabs®, cat number M0289S). 1/10 of the total volume of Antarctic phosphatase buffer and 3µl of Antarctic Phosphatase were added to the digested vector. The sample was incubated for 1 hour at 37°C, before being heated to 70°C for 5 minutes.

2.4.10.9. – Ethanol Precipitation of DNA

3 times the volume of 100% ethanol and 1/10 volume of 3M sodium acetate was added to each sample. Samples were incubated at 4°C for 30 minutes. They were then centrifuged at 13, 000RPM at 4°C for 20 minutes. The supernatant was removed and the pellet was washed in 70% ethanol. The samples were centrifuged at 13,000RPM at 4°C for 10 minutes, and the ethanol wash was removed. The pellet was left to dry and the DNA was resuspended in 15µl ddH_2O.

2.4.10.10. – Gel Separation of Digest Products and DNA Extraction

The digested samples were then loaded on either an agarose gel or a LMP agarose gel packed in ice (Life Technologies™, cat number 16520-050), and electrophoresed. The gels were then imaged and the appropriate bands extracted. The DNA was recovered using a MinElute Gel Extraction Kit (Qiagen™, cat number 28604) or, for the last cloning step, a QIAEX II Gel Extraction Kit (Qiagen™, cat number 20021) and eluted in 10µl EB buffer, according to the manufacturer's instructions. The DNA was quantified using a NanoDrop™.

		Gel type	Gel percentage (%)	Voltage and run time	Bands extracted
1 – Cloning mCherry into the pTRE3G plasmid	mCherry-pMK-RQ	Agarose	2.0	110V 2 hrs	746 bp
	pTRE3G	Agarose	0.8	110V 2 hrs	Linearised pTRE3G
2 – Cloning the KS linker sequence into pBluescript KS	KS-pMA-RQ	Agarose	2.5	110V 2 hrs	148 bp
	pBluescript KS	Agarose	2.5	110V 2 hrs	Linearised pBluescript KS
3 – Cloning the AM linker into pCMV-Tag1	AM-pMA-T	LMP agarose	5.0	110V 1 hr	84 bp
	pCMV-Tag1	Agarose	1.0	110V 2 hrs	Linearised pCMV-Tag1
4 – Cloning eGFP into pBluescript KS-KS (product of step 2)	pCMV - eGFP	Agarose	2.5	110V 2.5 hrs	800 bp
	pBluescriptKS-KS	Agarose	2.5	110V 2.5 hrs	Linearised pBluescript KS-KS
5 – Cloning pTRE3GmCherry (product of step 1) into pCMV-AM (product of step 3)	pTRE3GmCherry	Agarose	1.0	110V 2 hrs	1.126 Kb
	pCMV-AM	Agarose	1.0	110V 2 hrs	Linearised pCMV-AM
6 – Cloning i4x5 into pBluescript KS-eGFP (product of step 4)	i4x5	Agarose	1.0	110V 2 hrs	800 bp
	pBluescriptKS-eGFP	Agarose	1.0	110V 2 hrs	Linearised pBluescript KS-eGFP
7 – Putting eGFP into frame (using the product of step 6)	pBluescript i4x5-eGFP	Agarose	1.0	80V 2hrs	Linearised pBluescript i4x5-eGFP
	pBluescript i4x5-eGFP	Agarose	1.0	80V 2 hrs	Linearised pBluescript i4x5-eGFP
8 – Cloning modified H13i3 into pMAT-KP	Modified H13i3	Agarose	1.5	100V 1 hr	550 bp
	pMAT-KP	Agarose	0.8	100V 1 hr	Linearised pMAT-KP

9 – Cloning i3x4i4 into pMAT-modi3 (product of step 8)	i3i4	Agarose	1.0	100V 2 hrs	2.4 Kb
	pMAT-modi3	Agarose	1.0	100V 2 hrs	Linearised pCMV-modi3pTRE3GmCherry
11 – Cloning H13 sequence into pCMV-modi3i3i4pTRE3GmCherry (product of step 10)	pMAT-H13 sequence	Agarose	0.8	100V 1hr	6 Kb
	pCMV-pTRE3GmCherry	Agarose	0.8	100V 1hr	Linearised pCMV-pTRE3GmCherry
12 – Cloning UCOE into pCMV-H13SeqpTRE3GmCherry (product of step 11)	UCOE (MA892)	LMP Agarose	1.0	100V 1hr	2.2 Kb
	pCMV–H13SeqpTRE3GmCherry	LMP Agarose	1.0	100V 1hr	Linearised pCMV-H13SeqpTRE3GmCherry
13 – Cloning i4x5eGFP into pCMV-UCOEH13SeqpTRE3GmCherry (product of step 12)	pKS-i4x5eGFP	Agarose	0.8	100V 1 hr	1.6 Kb
	pCMV-UCOEH13SeqpTRE3GmCherry	Agarose	0.8	100V 1 hr	Linearised pCMV-UCOEH13SeqpTRE3GmCherry
14 – Cloning poly(A) into pCMV-UCOEH13SeqpTRE3GmCherryi4x5eGFP (product of step 13)	pMAT-H13polyA	LMP Agarose	1.5	100V 1.5 hrs	600 bp
	pCMV-UCOEH13SeqpTRE3GmCherryi4x5eGFP	Agarose	0.8	100V 1.5 hrs	Linearised pCMV-UCOEH13SeqpTRE3GmCherryi4x5eGFP

Table 2.8 – Summary of type of gel and conditions used to separate out the fragments of interest from the digestion reaction.

2.4.10.11. – Ligation

According to the protocol in section 2.4.6.

2.4.10.12. – Transformation into Chemically Competent Cells

According to the protocol in section 2.4.3.7. with the following alterations.

For cloning step eight 10µl of the ligation reaction was added to 150µl competent *E.coli* cells, which were then added to 300µl of S.O.C. medium (Invitrogen, cat number 15544-034), following the protocol above. For cloning step 14 One Shot® Top10 chemically competent *E. coli* (Life Technologies, cat number C4040-10) were used.

Ampicillin resistant (100µg/ml)	Kanamycin resistant (50µg/ml)
pBluescriptKS	pCMV-Tag1
pTRE3G	pCMV-eGFP
AM-pMA-T	mCherry-pMK-RQ
KP-pMA-T	BH-pMK-RQ
KS-pMA-RQ	x1i3-pMK-RQ
H13i3 mod – pMA-T	
MA892	
H13polyA-pMA-T	

Table 2.9 – Table of antibiotic resistance for plasmids used to generate the constructs, and the concentrations the antibiotics were used at.

2.4.10.13. - DNA Extraction from Bacterial Cultures

According to the protocol in section 2.4.8.

2.4.10.14. – Sequencing

According to the protocol in section 2.4.9. For primer details see appendix 7.2.

2.4.10.15. – Generating Glycerol Stocks

3µl of a positive sample was then transformed into competent *E.coli* cells (as above), and a glycerol stock was produced for storage at -80˚C.

2.4.11. – Generation of Tetracycline-responsive Cell Lines

Using the Tet-On 3G Inducible Expression System (Clontech Laboratories Inc, cat number 631168).

2.4.11.1. – Transfection

200,000 3T3 or HEK 293 cells were seeded into 1 well of a 6 well plate in complete growth medium, and left overnight. Xfect polymer was thawed and vortexed. For each well 2µg pCMV-Tet3G was added to Xfect reaction buffer, to a total of 100µl, and vortexed for 5 seconds. 0.6µl Xfect polymer was added to the diluted DNA, and vortexed for 10 seconds, before being incubated at room temperature for 10 minutes. The entire 100.6µl mixture was added drop by drop to the well. The plate was rocked gently to mix, before being incubated at 37°C for 4hrs. The medium was then removed from the well and replaced with 2mls of complete growth medium.

48hrs later each well was split into four 10cm dishes. 48hrs later G418 sulphate (Geneticin® Selective Antibiotic; Life technologies, cat number 11811-023) was added to the medium at a concentration of 50µg/ml to select for cells that had taken up the pCMV-Tet3G plasmid. The medium was changed every four days, until G418 resistant colonies appeared, and were visible to the naked eye.

Cloning discs (Sigma Aldrich, cat number Z374431-100EA) were soaked in trypsin. The medium was removed from the 10cm dish. Sterile tweezers were used to place individual cloning discs over each discrete colony. The dish was incubated at 37^{o}C for a few minutes, until cells had detached from the plate and attached onto the disc. Sterile tweezers were used to move each cloning disc to a well of a 24 well plate, already

containing medium containing G418. Once cells had migrated off the cloning disc into the well, the disc was removed. When the well was confluent the cells were split and transferred into 3 wells of a 6 well plate for screening for their response to doxycycline using the Dual- luciferase reporter assay.

2.4.11.2. – Dual-luciferase Reporter Assay

Transfect as in the first paragraph of 2.4.11.1, but use 5µg of pTRE3G-Luc and 0.5µg Renilla, and 1.5µl Xfect Polymer per well. 4hrs after the DNA:Xfect polymer mix has been added to the cells, the medium is removed and replaced with 2mls of complete growth medium, either with or without 1µg/ml doxycycline. After 24hrs the cells are washed once in 1xPLB (Passive Lysis Buffer). 500µl of 1xPLB is added to each well, and the plates were rocked gently at room temperature for 15 minutes. The cells were transferred to 1.5ml microfuge tubes and stored at -80^{o}C for at least 24 hrs. Once the lysates were thawed 20µl was added to three wells of a 96 well plate. The plate was loaded onto a single injector luminometer. 50µl of LARII (luciferase assay substrate re-suspended in luciferase assay buffer) was added to each well and the firefly luciferase activity was measured. 50µl of Stop and Glo reagent (50x Stop and Glo substrate diluted in Stop and Glo buffer) was added to each well, and the Renilla luciferase activity was measured. The Dual-luciferase reporter assay was performed using the kit from Promega, cat no. E1910.

2.4.11.3. – Transfection of pCMV-A and pCMV-B into HEK 293 Cells

Transfect as in the first paragraph of 2.4.11.1, but use 2µg of pCMV-A or pCMV-B, and 0.6µl Xfect Polymer per well. 4hrs after the DNA:Xfect polymer mix has been added to the cells, the medium is removed and replaced with 2mls of complete growth

medium, either with or without 1µg/ml doxycycline. After 24hrs the cells are washed once in PBS and collected for RNA and DNA extraction.

2.4.11.4. – DNA Extraction from Transfected Cells

A DNeasy Blood and Tissue Kit (Qiagen, cat number 69504) was used according to manufacturers instructions. The cells were washed in PBS. The pellet was resuspended in 200µl of PBS. 20µl of proteinase K and 4µl of RNase A were added to the sample, vortexed to mix, and then incubated at room temperature for 2 minutes. 200µl of Buffer AL was added to the sample, vortexed to mix, and then incubated at 56^{o}C for 10 minutes. 200µl of ethanol was added to the sample, and vortexed to mix, before the entire sample was transferred into a DNeasy Mini spin column. The sample was centrifuged at 6,000g for 1 minute, and the flow-through was discarded. 500µl of Buffer AW1 was added to the sample. The sample was centrifuged at 6000g for 1 minute, and the flow-through was discarded. 500µl of Buffer AW2 was added to the sample. The sample was centrifuged at 6,000g for 3 minutes, and the flow-through was discarded. The DNeasy Mini spin column was transferred to a fresh 1.5ml microfuge tube. 200µl of Buffer AE was added directly to the membrane of the column, and incubated at room temperature for 1 minute. The sample was centrifuged at 6,000g for 1 minute to elute the DNA. The DNA was quantified using a NanoDrop™ and stored at -20^{o}C.

2.4.11.5. – Quantitative Real-time PCR

According to the protocol in section 2.3.5. For primer details see appendix 7.2.

2.5. – Acknowledgement of Equipment Funding

Generation of the ChIP libraries and their sequencing required the use of the BRC Core Facilities provided with financial support form the Department of Health through the National Institute for Health Research (NIHR) comprehensive Biomedical Research Centre award to Guy's and St Thomas' NHS Foundation Trust in partnership with King's College London and King's College Hospital NHS Foundation Trust.

Chapter 3

Genomewide Identification of Co-localisation Sites for CTCF, cohesin, ATRX and MeCP2

3.1. - Introduction

Previous studies from Professor Oakey's laboratory have used a ChIP-Seq (chromatin immunoprecipitation followed by high-throughput sequencing) approach to map genomewide the occupancy of protein binding sites by the insulator protein CTCF, and by cohesin, the protein complex associated with both cell division and transcriptional regulation [22]. ChIP-Seq was performed on postnatal day 21 BxC and CxB mouse brain tissue, using antibodies specific for CTCF and the Rad21 subunit of cohesin. CTCF and cohesin were shown to bind together at around 27,000 sites across the genome, as well as independently at just over 22,000 sites for CTCF and 25,000 sites for cohesin. This is consistent with previous studies showing both independent and co-ordinated roles for these two factors. We found that the majority of DMR's are bound by CTCF, or cohesin, or both CTCF and cohesin, and that CTCF only binds in an allele-specific manner at or near imprinted loci, but not at all imprinted loci. No allele-specific binding was detected for cohesin, although when cohesin was bound with CTCF when CTCF was bound in an allele-specific manner cohesin tended to bind on the same allele [22].

gDMR Information (WAMIDEX)			CTCF and cohesin Binding Determined by ChIP-Seq			
gDMR	gDMR Position	Methylated Allele	CTCF		Cohesin	
			Binds	Allele	Binds	Allele
Bound by CTCF and cohesin precisely colocalised at gDMR						
GRB10	Chr11: 12, 025, 482 – 12, 025, 787	M	Yes	-	Yes	-
H19/IGF2	Chr7: 142, 580, 263 – 142, 582, 519	P	Yes	M	Yes	M
INPP5F_V2	Chr7: 128, 688, 274 – 128, 688, 642	M	Yes	Bi	Yes	Bi
Mcts2	Chr2:152, 686, 755 -152, 687, 275	M	Yes	-	Yes	-
MEST	Chr6: 30, 736, 488 – 30, 739, 335	M	Yes	P	Yes	P
NNAT	Chr2: 157, 560, 050 – 157, 561, 662	M	Yes	-	Yes	-
PEG13	Chr15: 72, 806, 335 – 72, 811, 649	M	Yes	P	Yes	P
Plagl1 (Zac1)	Chr10: 13, 090, 470 – 13, 090, 798	M	Yes	Bi	Yes	Bi
Bound by CTCF and cohesin colocalised at gDMR						
Cdh15	Chr8: 125, 387, 861 – 125, 390, 344	M	Yes	P	Yes	-
NESPAS	Chr2: 174, 295, 707 – 174, 300, 981	M	Yes	-	Yes	-
Zrsr1	Chr11: 22, 971, 842 – 22, 972, 319	M	Yes	-	Yes	Bi
ZIM2 (Peg3)	Chr7: 6, 727, 576 – 6, 732, 116	M	Yes	P	Yes	P
Bound by CTCF only						
PEG10	Chr6: 4, 747, 209 – 4, 747, 507	M	Yes	-	No	-
Meg3/Dlk1	Chr12: 109, 523, 353 – 109, 530, 779	P	Yes	-	No	-
IMPACT	Chr18: 12, 972, 197 – 12, 973, 741	M	Yes	-	No	-
Bound by cohesin only						
IGF2R/AIR	Chr17: 12, 741,297 – 12, 742, 707	M	No	-	Yes	-
GNAS-EXON1A	Chr2: 174, 326, 930 – 174, 329, 007	M	No	-	Yes	-
KCNQ1OT1	Chr7: 143, 295, 155 – 143, 295, 492	M	No	-	Yes	-
SNURF/SNRPN	Chr7: 60, 004, 992 – 60, 005, 415	M	No	-	Yes	-
No binding						
Nap1l5	Chr6: 58, 906, 696 – 58, 907, 062	M	No	-	No	-
Rasgrf1	Chr9: 89, 879, 568 – 89, 879, 853	P	No	-	No	-
Slc38a4?	Chr15: 96, 885, 270 – 96, 886, 284	M	No	-	No	-

Table 3.1 – Summary of the binding of CTCF and cohesin to gDMR's in the mouse. M = maternal, P = paternal, Bi = biallelic. Adapted from Prickett *et al*, 2013 [22].

It has been shown that ATRX (a chromatin remodeler) and MeCP2 (a methylation-sensitive DNA binding protein) bind with CTCF and cohesin at many loci across the genome, including at the *H19* ICR and the *Gtl2/Dlk1* imprinted regions [147]. It is currently unknown whether these 'super complexes' provide a general mechanism for

the regulation of gene expression, or whether they play a role specific to the regulation of gene expression from imprinted loci. By investigating the co-occupancy and potential interaction of these four proteins genomewide we hope to provide evidence that will lend support to one of these hypotheses, improving our understanding of the mechanisms underlying this type of regulatory process.

At least two alternative models have been proposed to explain how the interaction of ATRX, MeCP2, CTCF and cohesin at a single genomic location can regulate transcription. These are illustrated in Figure 3.1, using the *H19* DMR as an example. This locus has been studied in depth for CTCF binding [24] and gene expression [23] and so provides an excellent model to draw from. In both proposed models, CTCF is bound to the un-methylated maternal allele, and interacts directly with cohesin through the SCC3 subunit [150]. In the 'Imprinting super-complex' model (Figure 3.1) ATRX and MeCP2 interact with this complex on the maternal un-methylated allele. It is proposed that ATRX interacts with the SMC1 subunit of cohesin [147]. The position of MeCP2 is as yet unclear in the literature, although it is likely to interact with ATRX, since these proteins have been shown to interact in other studies [71]. In the 'Differential binding' model (Figure 3.1) ATRX and MeCP2 localise to the methylated paternal allele, away from the CTCF-cohesin complex on the maternal allele. This model takes into account MeCP2's preference for binding to methylated regions of DNA [98]; this arguably gives credence to the 'Differential binding' model since this preference is ignored by the 'Imprinting Super-Complex' model (which has MeCP2 associating with the un-methylated allele).

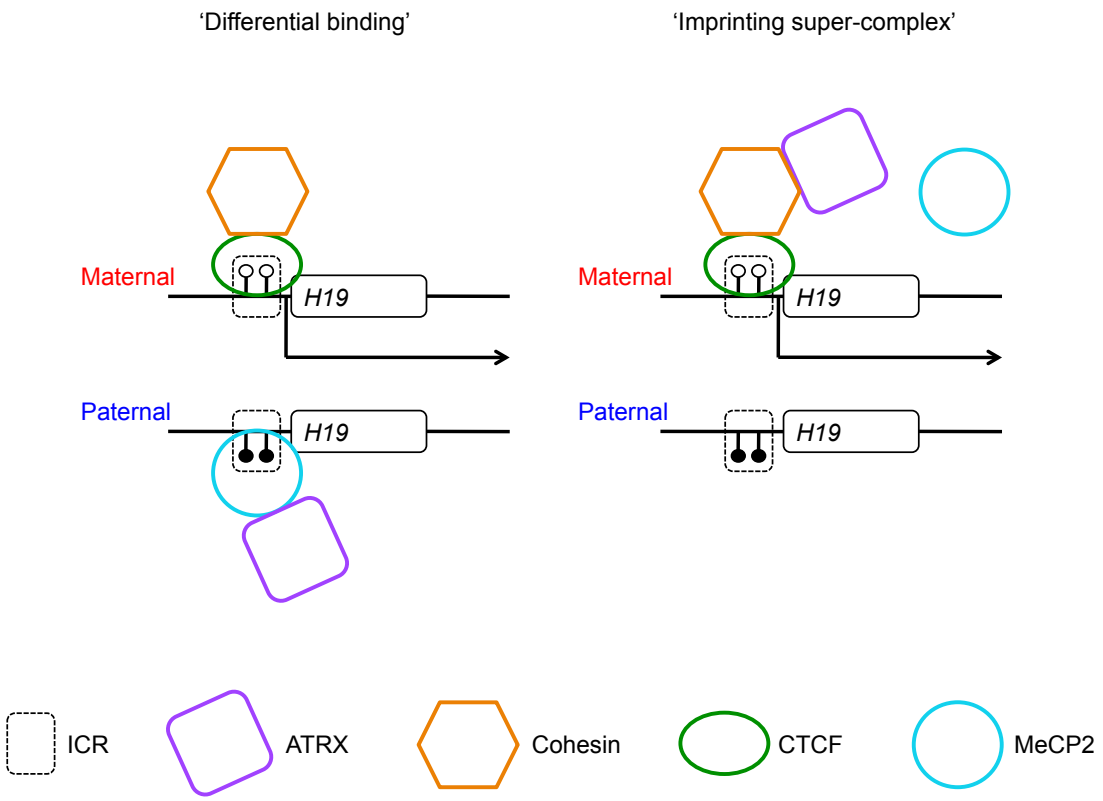


Figure 3.1 – Potential models of interaction between CTCF, cohesin, ATRX and MeCP2, shown at the *H19* DMR. The paternal allele is methylated (black circles) and the maternal allele is un-methylated (white circles). The arrows show the direction of transcription (reproduced from McCole, 2010 [151]).

We are therefore interested in clarifying how MeCP2 and/or ATRX interact with CTCF and cohesin. It is known that the four proteins form complexes at some imprinted loci [147], and by surveying genomewide we aim to determine whether this is a more widely adopted mode of gene regulation that occurs at other imprinted loci. We would also like to assay co-occupancy of these four proteins to find out if their interaction is limited to imprinted gene loci or if it occurs more widely across the genome. To determine this we undertook ChIP-Seq experiments using antibodies specific for ATRX and MeCP2, in order to locate these proteins on the genome. We then compared these results with those generated from ChIP-Seq experiments performed on CTCF and cohesin. Positive results from a ChIP-Seq experiment are referred to as 'peaks'; these

are genomic regions where sequencing reads for the protein of interest pile up on top of one another, implying a binding site for the protein.

To allow us to determine the interactions of ATRX, MeCP2, CTCF and cohesin in the context of imprinted genes we used tissues from intercross mice. These are the offspring of parents from different inbred sub-species of mice; genomic sequence differences between the two sub-species thus make it possible to determine the parent of origin of each allele of an imprinted gene (Figure 3.2). In this work we utilised mice generated from crosses between *Mus musculus castaneus* and C57BL/6 (both with *Mus musculus castaneus* as the mother and C57BL/6 as the father, and the reciprocal cross). Single nucleotide polymorphisms (SNPs) (a difference in the base present at a particular location in the sequence between two samples) occur roughly every 200-300bp between these two sub-species, enabling us to determine the parental inheritance of each allele of an imprinted gene. We were therefore able to investigate differential binding patterns for our proteins of interest between the maternal and paternal alleles of imprinted genes in these mice.

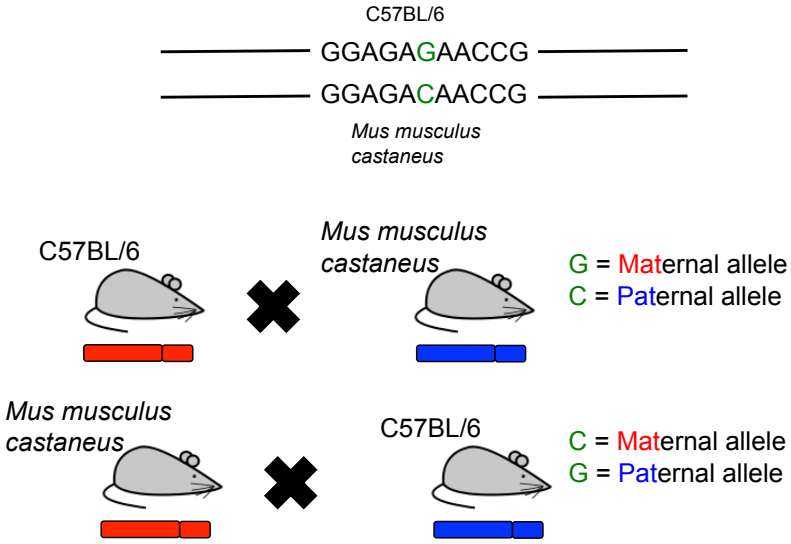


Figure 3.2 - Summary of the assignment of parental inheritance using SNPs. The SNP in the sequence is in green. In the C57BL/6 genome this is a G, in the *Mus musculus castaneus* genome this is a C. By comparing the sequence generated in the sequencing experiments to both parental genomes, parental inheritance of this region of sequence can be determined.

We used brain tissue as most imprinted genes are expressed in the brain. The CTCF and cohesin ChIP-Seq libraries were generated from intercross brain tissue and so using this for the ATRX and MeCP2 ChIP-Seq will allow us to compare between all four datasets easily.

3.2. – Results

3.2.1. - Optimisation of Chromatin Immunoprecipitation

ChIP-Seq was previously performed in Professor Oakey's laboratory for CTCF and cohesin (using antibodies to the Rad21 subunit). Our initial approach to ChIP-Seq using ATRX and MeCP2 antibodies therefore followed the same protocol [22] (protocol 1; see Materials and Methods). As our aim is to compare binding sites for all four proteins, this approach has the advantage that all four datasets would be generated using the same tissue type. In protocol 1, ChIP-Seq is performed using CTCF, ATRX and MeCP2 antibodies on BxC and CxB brain tissue from p21 mice. This protocol was

attempted 20 times using a range of concentrations of the ATRX and MeCP2 antibodies, and both protein A and G agarose beads. In each case we were unable to extract sufficient chromatin to generate sequencing libraries. The library preparation kit that we used for the CTCF and cohesin ChIP-Seq (NEBNext ChIP-Seq Library Prep Master Mix Set for Illumina (NEB, cat no E6240)) requires 10ng of ChIP DNA for each library. Performing ChIP on multiple samples for both ATRX and MeCP2 and pooling them didn't yield enough DNA to take forward for library preparation. Previous work in the laboratory had shown that to generate the optimal quality libraries ChIP DNA needs to be processed within two weeks of its generation, making it difficult to pool samples from consecutive ChIP experiments, check the efficiency using qPCR and check the fragment size and re-sonicate if necessary all within this time frame.

The generation of sequencing libraries is a multi-step process, where chromatin can be lost at every step. Sufficient chromatin needs to be extracted from the ChIP for qPCR validation, the quantification of the DNA concentration, the analysis of fragment sizes and qPCR quantification of the libraries produced. In the literature it is clear that most successful applications of ChIP-Seq have been performed in cell lines, rather than tissues. As such there are relatively few published or commercially available tissue-based ChIP-Seq protocols.

One commercially available protocol optimised for use in cell lines is EZ-ChIP^{TM} kit (ChIP protocol 2; see Materials and Methods). Although our experiments are tissue-based, we were encouraged to employ this kit due to its successful use for ATRX and MeCP2 ChIP by Kernohan *et al* [147], albeit in cell lines. However after four trials we were unable again to extract enough chromatin for the preparation of libraries. This is

most likely due to the fact that we performed the experiments in whole tissue samples. In particular, possible reasons for the failure of this protocol could include: incompatibility of one of the buffers or reagents we used for chromatin extraction with those used in the kit (which supplies its own reagents for chromatin extraction in cell lines); incomplete dissociation of tissue samples into individual cells due to tissue structure (thus inhibiting chromatin extraction or reducing the efficiency of chromatin extraction); or incompatibility of one of our antibodies with the reagents in the kit. Note that we were unable to perform this experiment in a cell line because the in-house BxC/CxB mouse intercross system that allows parental inheritance of alleles to be determined is tissue-based. One option could have been to breed intercross mice and derive cell lines, however by the time these problems came to light, our *Mus musculus castaneus* mice were having fertility problems in the animal facility and by the time these were resolved, there was not sufficient time to generate the lines and do the ChIP-Seq.

Finally, ChIP protocol 3 was designed after consultation with Professor Adrian Bird (Wellcome Trust Centre for Cell Biology, University of Edinburgh) and Professor Richard Gibbons (Weatherall Institute of Molecular Medicine, University of Oxford) whose laboratory groups have performed ChIP on MeCP2 and ATRX respectively. Professor Bird advised us that since MeCP2 binds to numerous CpG rich regions in the genome, it is typical to see high background levels of MeCP2 binding in ChIP-Seq experiments making the interpretation of this data at regions of interest extremely challenging. For this reason optimisation of the ATRX ChIP-Seq experiments was prioritised. ChIP protocol 3 was thus based on a protocol provided by Dr Hsiao Voon in Professor Gibbons' laboratory group for ChIP-Seq on ATRX (see Materials and

Methods section). Dr Voon's protocol is optimised for use with ES cells, and attempts to implement this protocol in tissue were unsuccessful.

Having been unable to successfully perform ChIP for our proteins of interest using protocols 1-3 in tissue, we considered performing ChIP in cultured cells as an alternative. This has the disadvantage that consistency with the existing CTCF and cohesin ChIP-Seq data sets, which were generated using BxC/CxB intercross tissue, is reduced; on the other hand we found that ChIP using cell lines could reliably generate sufficient DNA for sequencing library preparation. ChIP protocol 3 was performed in two cell lines. Firstly, Neuro2a cells, which are a *Mus musculus* neuroblast cell line were chosen because being a neural cell line they should be expected to be comparable with the CTCF and cohesin ChIP-Seq data sets that were generated in neural tissue. Secondly, we obtained reciprocal intercross embryonic stem (ES) cell lines from Dr Robert Feil (Institute Genetique Moleculaire Montpellier) [152]. His laboratory group use Japanese Fancy Mice (a species of *Mus musculus molossinus*) and C57BL/6 (a strain of *Mus musculus domesticus*), and have generated ES cells from the reciprocal crosses between these two inbred sub-species (JxB and BxJ). Like the *Mus musculus castaneus* and C57BL/6 reciprocal crosses, there are frequent SNPs between these strains of mice, which allow the parent of origin of each allele to be determined. In addition, the JF1 mouse genome has been sequenced, allowing us to perform sequencing experiments and compare the sequence to the reference genome. Since ChIP protocol 3 was successful in both cell lines, we decided to use the reciprocal intercross ES cells for subsequent analysis as they would allow the determination of parent-of-origin allele specific binding of ATRX and MeCP2, an advantage over the Neuro2a cells.

Protocol	Sample	Type of bead used	Variations	Outcome
1	BxC and CxB p21 brain	Protein A agarose	Concentration of antibodies Incubation time with antibodies qPCR primers	Little or no enrichment of MeCP2 and ATRX samples over the IgG control
		Protein G agarose	Concentration of antibodies Incubation time with antibodies qPCR primers	Little or no enrichment of MeCP2 and ATRX samples over the IgG control
2	BxC and CxB p21 brain	Protein G agarose	Incubation time with antibodies	No enrichment of MeCP2 and ATRX samples over the IgG control
3	BxC and CxB p21 brain	Dynabeads	Concentration of antibodies	Some enrichment of MeCP2 and ATRX samples over the IgG control, but can't be reliably replicated between experiments
	Neuro2a cells	Dynabeads	Concentration of antibodies Cell number	Good enrichment of MeCP2 and ATRX samples over the IgG control
	Intercross ES cells (JxB, BxJ)	Dynabeads	Concentration of antibodies Cell number	Good enrichment of MeCP2 and ATRX samples over the IgG control

Table 3.2 – Summary of ChIP protocols and conditions tried.

3.2.2. – Library Preparation for Sequencing

We generated libraries for an input sample (chromatin extracted and processed in the same way as the other chromatin samples, but not subject to ChIP), two ATRX samples for the JxB ES cells and a MeCP2 sample for both BxJ and JxB ES cells. We were unable to collect sufficient chromatin from the ATRX ChIP on the BxJ ES cells for generating sequencing libraries. Two libraries were included for ATRX for the JxB ES cells to ensure there was sufficient chromatin for sequencing, as the chromatin yields from these ChIP experiments were lower than for the MeCP2 ChIP experiments. This

was done using the DNA SMART™ ChIP-Seq Kit for Illumina. We used the Agilent 2200 Tapestation to assess the size of the fragments in each library generated (Figure 3.3). The average fragment length for each library is shown in table 3.3. A qPCR kit was used to quantify the amount of DNA in each library.

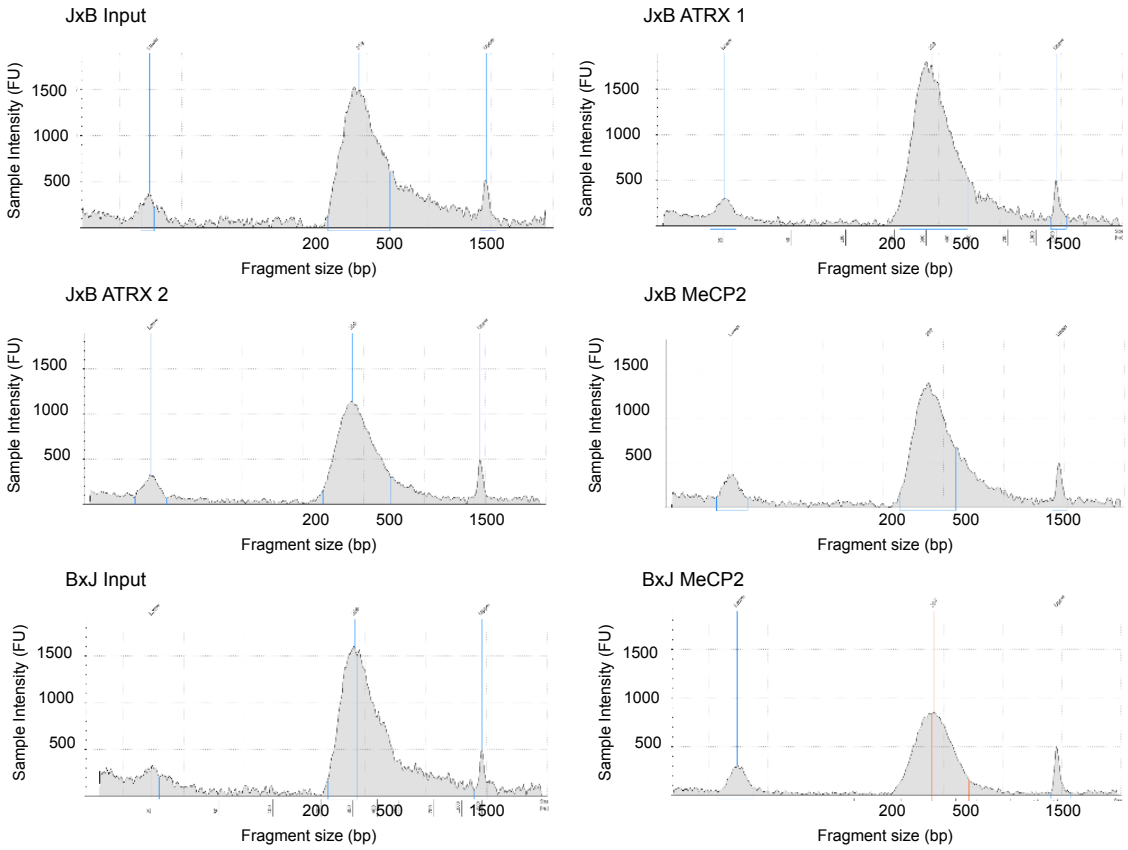


Figure 3.3 – Trace representation of fragment size analysis from the Agilent 2200 Tapestation for all six of the ChIP libraries prepared. All six of the libraries contain a peak between about 200 and 500bps, showing that the majority of fragments in these samples fall within this range.

Index no.	Sample	Average fragment size (bp)
2	JxB Input	339
3	JxB ATRX 1	336
4	JxB ATRX 2	332
5	JxB MeCP2	326
6	BxJ Input	338
7	BxJ MeCP2	321

Table 3.3 – Index used to label each library and the average fragment size of each library.

3.2.3. – Quality Control of ChIP-Seq Datasets and Read Statistics

The ChIP-Seq data for the JxB and BxJ MeCP2 experiments showed low-level binding across the genome, making it impossible to call specific peaks, even with the most lenient cutoffs. For this reason I have chosen to perform further analysis only on the two JxB ATRX datasets.

MACS2 was used to call peaks in the ATRX 1 and ATRX 2 ChIP-Seq datasets. Peaks are regions of the genome where multiple reads (sequences generated from the ChIP-Seq) align. As more reads are aligned to the genome, they start to stack on top of each other in regions where a large number of reads align. This is called a peak, and implies that the protein being investigated in the ChIP-Seq experiment, in this case ATRX, binds to this region.

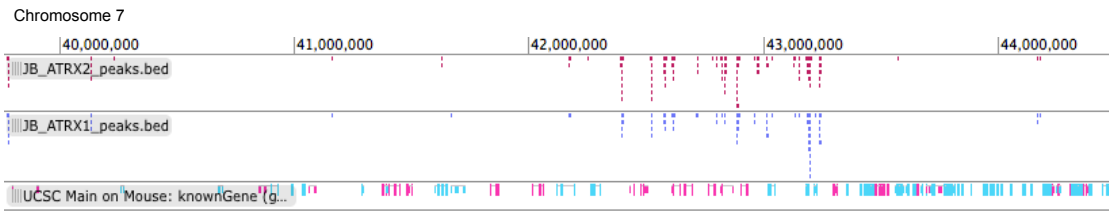


Figure 3.4 – Screenshot from Galaxy of the ATRX 1 and ATRX 2 reads aligning to form peaks along chromosome 7. The scale and genomic location at the top of the figure are from the UCSC mouse build mm10. The red dots are the JxB ATRX 2 reads, and the blue dots are the JxB ATRX 1 reads.

The quality of the reads generated was assessed in two ways. Phred scaled quality scores were generated for each base in the read. This is a measure of the accuracy of the assignment of each base in the read, and operates on a logarithmic scale. A score of 10 indicates a base call accuracy of 90%, whereas a score of 40 indicates a higher base call accuracy of 99.99% [153]. For JxB ATRX 1 the majority of bases have Phred scores of above 30 (see Figure 3.5 A), showing that the reads are of a good quality. JxB

ATRX 2 on the other hand, has Phred scores between 10 and 20 for the majority of bases in its reads, implying that this sequencing is of a poorer quality (see Figure 3.5 B).

The percentage of each base incorporated along the reads was also assessed. The four bases should be incorporated in similar proportions (roughly about 25% for each). Preferential inclusion of one base over the others can be an indicator of a problem with the sequencing reaction, the ChIP experiment itself or reading the sequence into the adaptors used to prepare the sequencing libraries. Identifying the presence of preferential inclusion of bases allows the reads to be trimmed to remove these portions of the sequences, improving the quality of the reads. For JxB ATRX 1 the base incorporation is as expected, except for the first couple of bases, this indicates that the reads are of a good quality (Figure 3.5 C). For JxB ATRX 2 the base incorporation is also as expected, except for the first 14-16 bases (Figure 3.5 D), indicating that the reads are of a good quality after this point. The reads for both ATRX 1 and 2 were trimmed to remove the problematic bases at the start.

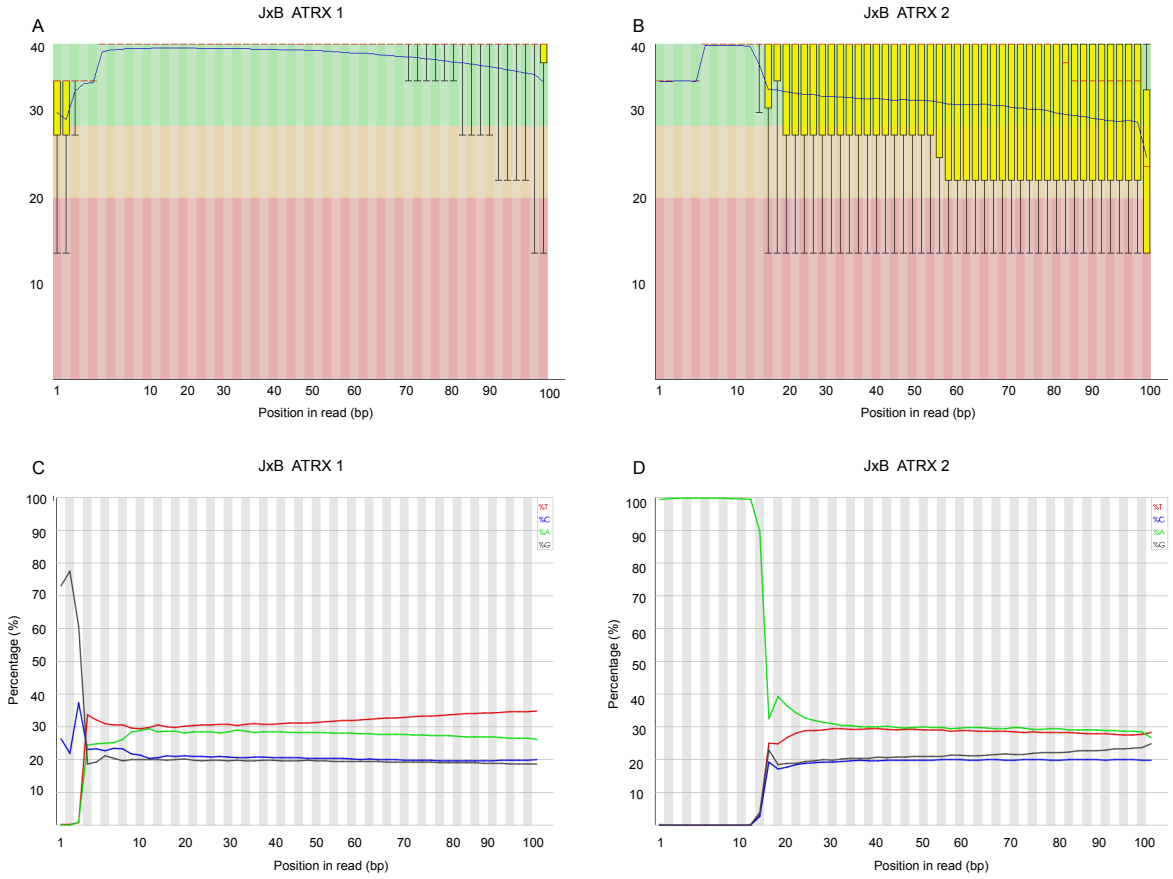


Figure 3.5 – Quality plots for the forward reads of the JxB ATRX 1 and 2 ChIP-Seq libraries. A and B - Phred scaled quality score box plot as a function of read position. This suggests a good quality for ATRX 1 (A), but a poor read quality for ATRX 2 (B). Both graphs show the expected drop in quality towards the 3' end. C and D - Base incorporation percentage as a function of read position. The uneven base incorporation in the first couple of bps for ATRX 1 (C) and the first 14-16 bps for ATRX 2 (D) shows that there has been a problem with the sequencing. The relatively flat lines for each of the bases beyond these points shows there has been little cycle-to-cycle variation. In C increasing C and T percentages at the 3' end could suggest reading into the adaptors.

37, 439, 989 reads were found in the JxB ATRX 1 dataset, whereas 62, 567, 501 reads were identified in the JxB ATRX 2 dataset. Of these reads 65.07% and 65.46% aligned to repetitive regions, in the JxB ATRX 1 and JxB ATRX 2 replicates respectively. About 15% of the reads generated (14.78% of the ATRX 1 reads, and 15.08% of the ATRX 2 reads) didn't align to the genome. There were an average of 248 and 383 reads in each peak in the JxB ATRX 1 and JxB ATRX 2 libraries respectively. The difference in the number of ATRX reads between the two replicates is likely due to the difference in the quality of the samples. It is possible that the inaccuracies in assigning

bases in the ATRX 2 reads may have caused them to align to other regions of the genome, as well as regions where ATRX binds, artificially increasing the read count.

The cluster density for the libraries was 763 +/- 75 k/mm^2. The libraries were barcoded and run on the same cell of a flow cell. This was a little low compared to other cells on the same flow cell, with six of the seven other cells having a cluster density of over 800 k/mm^2 and two being over 1000 k/mm^2. We generated libraries containing long reads as we wanted to be able to align the reads to the parental genomes, using SNP's between the two, to allow us to assign parental inheritance and look for strain specific differences.

ATRX ChIP-Seq data generated by Dr Gibbons' laboratory on mouse ES cells (a published dataset in the public domain) contained 11,638,285 reads after the removal of duplicate reads [154]. Our read counts are much higher than this. Neither of our ATRX sequencing libraries is of as high a quality as we would like, and this could explain why we appear to have gained reads.

3.2.4. – ATRX Binding Genomewide

ATRX binds at many different locations across the mouse genome. We identified 12,925 peaks across the genome where ATRX binds. Peaks were defined as those significant where $q<0.01$. This is a much higher number than that identified in the ATRX ChIP-Seq data from Dr Gibbons' laboratory, which identified 1,305 binding sites [154]. This discrepancy is probably due to the inaccuracies in assigning bases in our reads causing them to align to other regions of the genome, as well as regions where ATRX binds, artificially increasing the number of binding sites identified.

3.2.5. – Validation of ATRX and MeCP2 Chromatin Immunoprecipitation

To validate the results obtained via ChIP-Seq we performed qPCR on the chromatin generated from both the ATRX and MeCP2 ChIP experiments, and their matched inputs (chromatin which was processed through the ChIP protocol but which wasn't exposed to an antibody). A standard curve consisting of five different concentrations of DNA was also run for each primer set, allowing the ChIP qPCR results to be quantified. The results of the ATRX and MeCP2 qPCR's are shown as a percentage of the qPCR results of the input for the same primer set. This analysis strategy was used by Kernohan *et al* [147], and is more appropriate than running an endogenous control, which could be bound to different extents by the antibodies used. We confirmed that ATRX binds within the ICR of H19 and to Nap1l5, but not within the promoter of GAPDH or MNT (Figure 3.6 A and B), in JxB ES cells (one Figure per ATRX input sample). qPCR for MeCP2 confirmed binding within the ICR of H19 and Gtlk-GD3 (part of the Gtl2/Dlk1 imprinted region defined by Kernohan *et al* [147]), but not a CpG negative region downstream of GAPDH on chromosome 14 (labelled here as CpG negative), in both JxB (Figure 3.6 C) and BxJ (Figure 3.6 D) ES cells. The difference in binding at *Gtlk-GD3* between the JxB MeCP2 ChIP compared to the BxJ MeCP2 ChIP could be due to the presence of a SNP between the C57BL/6 and Japanese Fancy Mouse genomes causing a difference in the affinity of MeCP2 to one parental genome over the over at this locus. The locations of the regions amplified in the qPCR experiments are shown in Figure 3.7.

ChIP was performed on BxC p21 mouse brain using the CTCF antibody as a positive control for the ChIP protocols. Previous work in the laboratory has shown that CTCF

reliably binds to intron 10 of H13 and doesn't bind at intron 3 of H13. I saw the same results with my ChIP with CTCF (Figure 3.6 E).

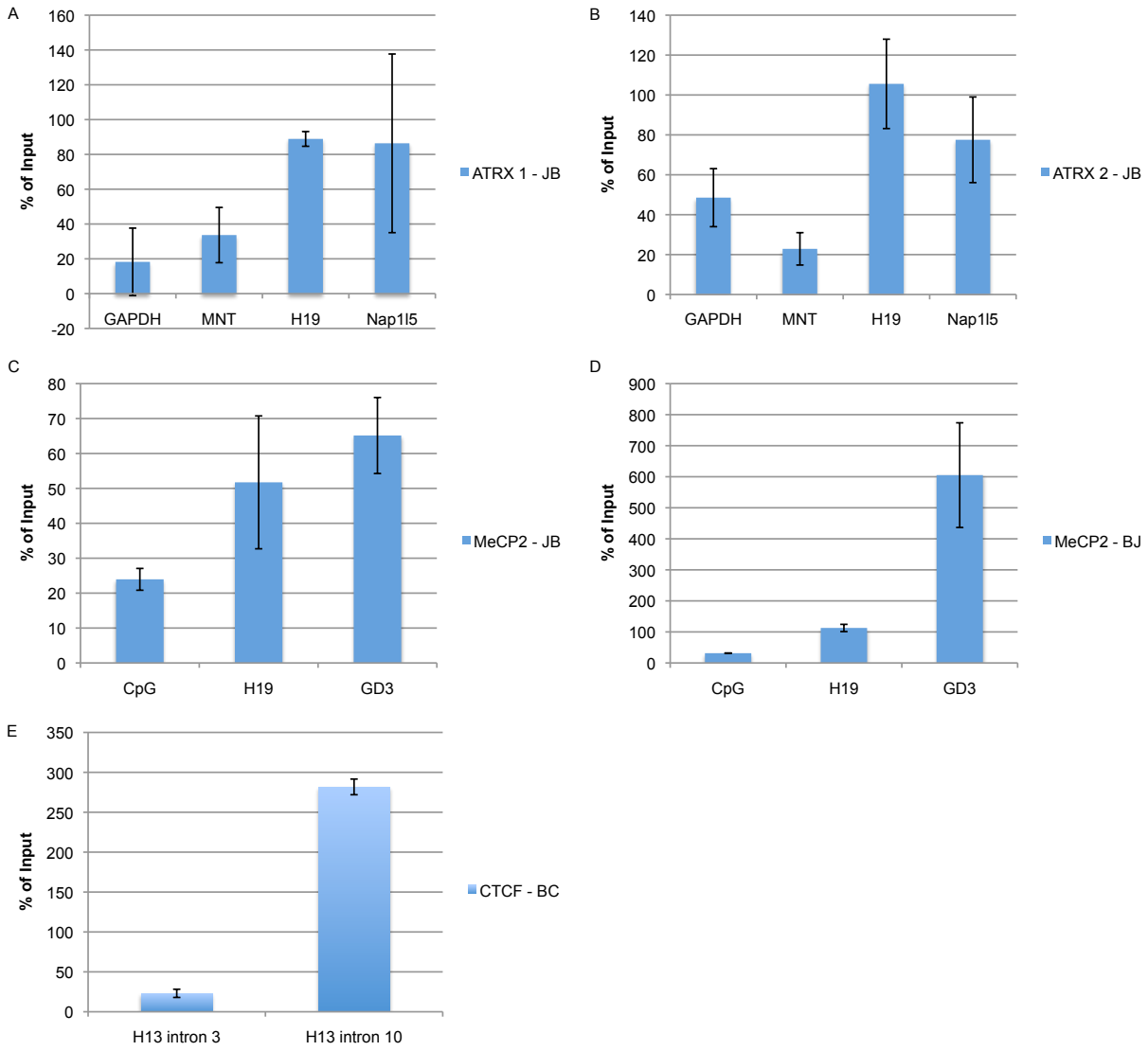


Figure 3.6 – qPCR validation of ChIP. The error bars are the standard deviation of the sample. For A and B *GAPDH* and *MNT* were regions where ATRX doesn't bind, and *H19* and *Nap1l5* were regions where ATRX does bind. For C and D a CpG negative region, just downstream of *GAPDH* on chromosome 14, is one region where MeCP2 doesn't bind, and *H19* and *GTLK-GD3* were regions where MeCP2 does bind. For E part of intron 3 of *H13* acts a negative control region for CTCF, and part of intron 10 of *H13* acts as a positive control, where CTCF does bind. n=3 technical replicates

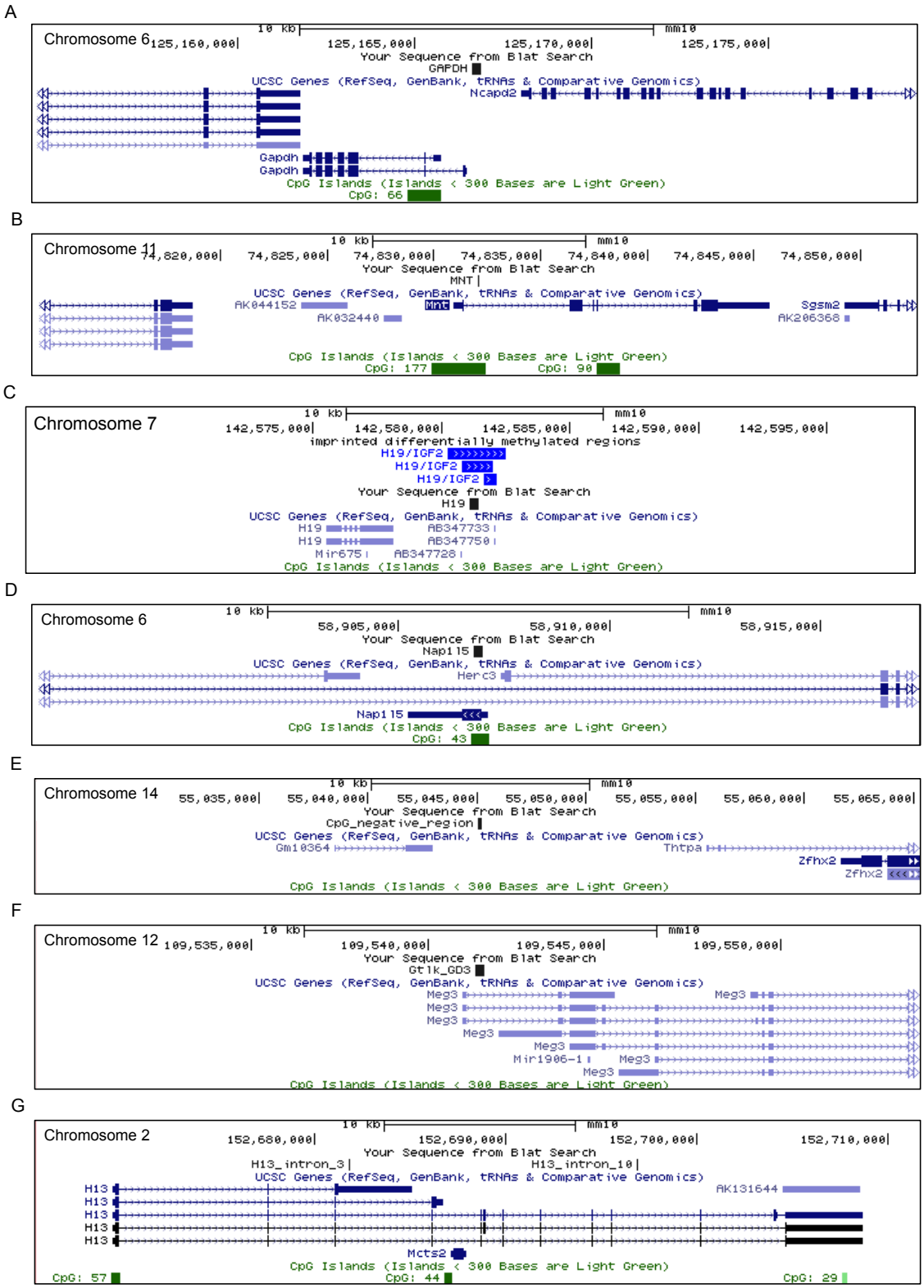


Figure 3.7 – The genomic locations of the regions amplified by qPCR. A and B show the locations of regions where ATRX does not bind (at the promoter of *GAPDH* and within intron 1 of *MNT* respectively). C and D show the location of regions where ATRX does bind (within the ICR of *H19* and *Nap1l5* respectively). E shows the location of a CpG negative region where MeCP2 does not bind. C and F show the locations of regions where MeCP2 does bind (within the ICR of *H19* and within an intron of *Meg3*, part of the *Gtl2/Dlk1* imprinted region, respectively). G shows the locations of a region not bound by CTCF (within intron 3 of *H13*), and a region bound by CTCF (within intron 10 of *H13*). Taken from the UCSC genome browser [155], using the mouse GRCm38/mm10 reference genome (published in 2011.)

3.2.6. –Allele-specific Binding of ATRX at Imprinted Regions

ATRX was shown to bind at three of the 22 imprinted gDMR's studied. Interestingly the only gDMR that is bound by ATRX, CTCF and cohesin is *Mcts2*, which is the imprinted retrogene in our model imprinted locus studied in Chapter 4. ATRX binds to the paternal allele of *Mcts2*, but allele-specific binding could not be determined for CTCF and cohesin (Figure 3.8 A). In *Mcts2* the paternal allele is un-methylated. The binding of ATRX to the un-methylated allele supports the 'Imprinting super-complex' model, in which ATRX, MeCP2, CTCF and cohesin all bind to the same un-methylated allele. The allele-specific binding of MeCP2, CTCF and cohesin needs to be determined at this locus to inform on which model more accurately represents the interaction of these four proteins.

ATRX is shown to bind at the *GTL2/DLK1* iDMR (on chromosome 12; 109, 523, 353 – 109, 530, 779) (Figure 3.8 B), along with CTCF, and at *IGF2R/AIR* (on chromosome 7; 142, 580, 263 – 142, 582, 519) (Figure 3.8 C) with cohesin. It is likely that ATRX binds to more imprinted regions, for example PCR on ATRX ChIP (using murine brain) has shown that ATRX binds to the DMR of *H19* [147], which we are unable to see in our sequencing data. Improved library quality or read depth would resolve this problem.

gDMR Information (WAMIDEX)			ATRX, CTCF and cohesin Binding Determined by ChIP-Seq					
gDMR	gDMR Position	Methylated Allele	ATRX		CTCF		Cohesin	
			Binds	Allele	Binds	Allele	Binds	Allele
Bound by ATRX, CTCF and cohesin								
Mcts2	Chr2:152, 686, 755 -152, 687, 275	M	Yes	P	Yes	-	Yes	-
Bound by ATRX and CTCF								
GTL2/DLK1	Chr12: 109, 523, 353 – 109, 530, 779	P	Yes	-	Yes	-	No	-
Bound by ATRX and cohesin								
IGF2R/AIR	Chr17: 12, 741,297 – 12, 742, 707	M	Yes	-	No	-	Yes	-
Bound by CTCF and cohesin								
GRB10	Chr11: 12, 025, 482 – 12, 025, 787	M	No	-	Yes	-	Yes	-
H19/IGF2	Chr7: 142, 580, 263 – 142, 582, 519	P	No	-	Yes	M	Yes	M
INPP5F_V2	Chr7: 128, 688, 274 – 128, 688, 642	M	No	-	Yes	Bi	Yes	Bi
MEST	Chr6: 30, 736, 488 – 30, 739, 335	M	No	-	Yes	P	Yes	P
NESPAS	Chr2: 174, 295, 707 – 174, 300, 981	M	No	-	Yes	-	Yes	-
NNAT	Chr2: 157, 560, 050 – 157, 561, 662	M	No	-	Yes	-	Yes	-
PEG13	Chr15: 72, 806, 335 – 72, 811, 649	M	No	-	Yes	P	Yes	P
U2AF1-RS1	Chr11: 22, 971, 842 – 22, 972, 319	M	No	-	Yes	-	Yes	Bi
ZAC1	Chr10: 13, 090, 470 – 13, 090, 798	M	No	-	Yes	Bi	Yes	Bi
ZIM2	Chr7: 6, 727, 576 – 6, 732, 116	M	No	-	Yes	P	Yes	P
Bound by CTCF								
IMPACT	Chr18: 12, 972, 197 – 12, 973, 741	M	No	-	Yes	-	No	-
PEG10	Chr6: 4, 747, 209 – 4, 747, 507	M	No	-	Yes	-	No	-
Bound by cohesin								
GNAS-EXON1A	Chr2: 174, 326, 930 – 174, 329, 007	M	No	-	No	-	Yes	-
KCNQ1OT1	Chr7: 143, 295, 155 – 143, 295, 492	M	No	-	No	-	Yes	-
SNURF/SNRPN	Chr7: 60, 004, 992 – 60, 005, 415	M	No	-	No	-	Yes	-

Table 3.4 – Summary of the binding of ATRX, CTCF and cohesin to gDMR's in the mouse. M = maternal, P = paternal, Bi = biallelic.

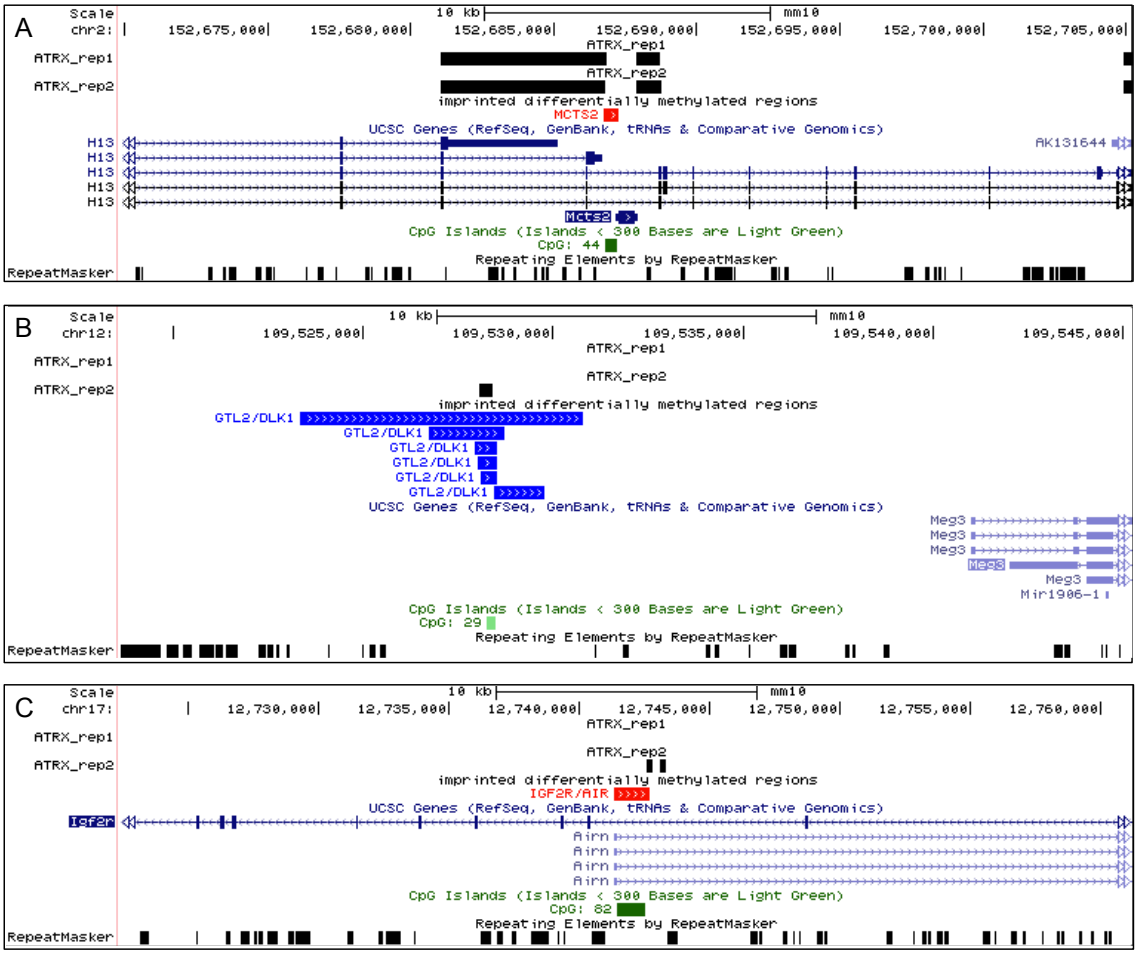


Figure 3.8 – Peaks of ATRX binding over *Mcts2* (panel A), *GTL2/DLK1* (panel B) and *IGF2R/AIR* (panel C). CpG islands are shown in green. Imprinted DMR's are shown in red (maternally methylated) and blue (paternally methylated). Taken from the UCSC genome browser [155], using the mouse GRCm38/mm10 reference genome (published in 2011.)

3.3. – Discussion

In this chapter we have discussed the optimisation of ChIP with ATRX and MeCP2 antibodies, the analysis of the data generated and comparison of this data with ChIP-Seq datasets generated previously in Professor Oakey's laboratory.

The ATRX and MeCP2 ChIP experiments required a lot of optimisation. It was difficult to find antibodies that had been previously used in ChIP-Seq experiments on tissue samples, rather than cell lines. There are more ATRX antibodies available now, so it may be useful to test some of these on our tissue samples. Tissues are difficult to

work with as they require the structure of the tissue to be dissembled completely to release individual cells in the final sample. If clumps of cells remain in the sample it can make it difficult for the antibody to bind properly causing binding sites to be missed, as access is blocked by the presence of other cells. It might be worth considering generating antibodies for ATRX and MeCP2 ourselves, or over-expressing a tagged version of the proteins and using an antibody specific for the tag.

There were several issues associated with the bioinformatics analysis of the ATRX and MeCP2 datasets, which need to be considered. About 15% of the reads in the ATRX 1 and 2 replicates didn't align to the genome. We didn't investigate these reads any further. The forward and reverse reads in all the datasets were considered independently, rather than being analysed as one complete set of reads, which may have affected the alignment of the reads to the genome. The SNP annotation used did not exactly match the mouse strain. A difference between the SNPs present in the SNP annotation used and the mouse strain may have stopped reads aligning properly to the genome, causing this information to be lost from the analysis. Another important point to consider is that we only had data from one cross, and not the reciprocal. Without the reciprocal cross we cannot rule out the possibility that the allele-specific peaks we found are due to a bias in the alignment of sequences to one genome over the other, rather than an actual preference for one parental allele over the other.

The apparent lack of overlap between the binding of ATRX, CTCF and cohesin at imprinted regions could be due to differential binding in the samples used. The CTCF and cohesin ChIP-Seq experiments used murine brain tissue, whereas the ATRX ChIP-Seq experiment utilised ES cells. It is possible that ATRX does co-localise to regions

where CTCF and cohesin bind but that these regions vary across tissue types. Ideally these three datasets would all have been generated from the same mouse tissue, using reciprocal crosses of the same two mouse strains, so that we could capture these interactions and interrogate them in an allele-specific manner.

Another possible explanation for the apparent lack of overlap in the binding sites of ATRX, CTCF and cohesin could be due to the quality of the sequences generated from the ATRX ChIP-Seq experiments. This data was not of an ideal quality, and we know from other published ATRX ChIP-Seq datasets that we are missing some ATRX binding sites. It is possible that by chance the binding sites we are missing are in the regions where CTCF and cohesin bind. If I had more time I would compare these published ATRX ChIP-Seq datasets to our CTCF and cohesin ChIP-Seq datasets to allow me to determine if this is the case. The published ATRX ChIP-Seq datasets are not generated from the same tissue types, and in some cases the same species, as our CTCF and cohesin ChIP-Seq datasets, but this would still give us an indication if the three proteins do bind to the same regions of the genome.

MeCP2 is a difficult protein to work with given its propensity to bind to methylated CpG's, of which there are many throughout the genome. This leads to low-levels of MeCP2 binding genomewide, making it very difficult to determine defined peaks at regions where it binds. A better way to investigate the importance of MeCP2 binding in collaboration with ATRX, CTCF and cohesin, and at imprinted regions might be to look at these on an individual basis. Once the regions where ATRX, CTCF and cohesin co-localise have been determined these can then be checked by ChIP and qPCR for binding

by MeCP2. This would reduce the background noise seen when investigating MeCP2 binding using high-throughput sequencing techniques.

The ATRX and MeCP2 ChIP-Seq data that we have generated from this project can not provide the data that we need determine the role of 'super-complexes' in the regulation of gene expression. We do not have allele-specific binding information for ATRX and MeCP2 so we are unable to discriminate between the two proposed models for 'super-complexes': the 'Imprinting Super-Complex' and the 'Differential binding' model. More work is needed before we can inform on these models. However, we have made a start, and more ChIP-Seq experiments in additional reciprocal crosses and deeper sequencing reads can resolve these issues.

Chapter 4

Examining the Mechanisms of Gene Regulation in the Context of a Well Studied Imprinted Gene Pair

4.1. – Introduction

We have investigated the mechanisms of gene regulation in the context of the *H13/Mcts2* imprinted locus. This locus was chosen because the presence of the imprinted retrogene (*Mcts2*) in one of the introns of the 'host' gene (*H13*) alters transcription through the host, imposing an allele-specific expression onto it, despite the fact that the promoter of the host gene itself is not imprinted (Figure 4.1). Imprinted retrogenes are a good model for examining mechanisms of gene regulation because even though the two alleles share an identical environment and are subject to the same influences, the gene is only expressed from one allele and not the other. Therefore this must be due to epigenetic factors operating *in cis*. Two approaches were used to investigate regulatory mechanisms at *H13*/*Mcts2*. Both approaches use purpose-built constructs to allow transcription through this locus to be controlled experimentally.

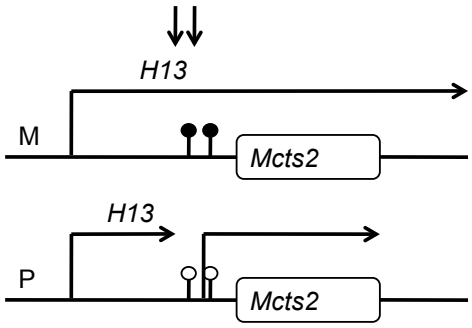


Figure 4.1 - Transcription of *Mcts2*, an imprinted retrogene, affects transcription of *H13*, the host gene. When *Mcts2* is not expressed, at the maternal allele (M), transcription occurs through *H13* to one of three downstream poly (A) sites. When *Mcts2* is expressed from the paternal allele (P), this causes premature termination of transcription of *H13*, and use of one of two upstream poly (A) sites. Transcripts are indicated by the horizontal arrows. The black circles show methylated CpG's and the white circles show un-methylated CpG's. The downward pointing arrows show the approximate locations of the two poly (A) sites located upstream of Mcts2. The other three poly (A) sites are located much further downstream of Mcts2, and are not shown in this diagram.

The first approach, the generation of *Mcts2* knock-in and knock-out mice, is part of a long-term project in Professor Oakey's laboratory, which aims to generate transgenic mice containing our constructs of interest, allowing transcription through the *H13/Mcts2* locus to be altered. The use of mice allows for the correct setting of methylation marks through meiosis at the site of the construct, which could be important for regulation of transcription at this site. During embryonic development, methylation is removed from the genome, thus removing any acquired epigenetic modifications. The genome is then re-methylated to reset essential methylation patterns [50].

Two constructs had previously been generated by Dr Mike Cowley, a post-doctoral researcher in the laboratory: one targeted to *H13* to knock out endogenous expression of *Mcts2* and the other targeted to *Fam13c* to insert *Mcts2* into one of its introns. *Fam13c* is located on chromosome 10 and has a similar structure to the *H13* locus, making use

of multiple poly (A) sites to generate transcripts of different sizes (Figure 4.2). However, *Fam13c* does not exhibit imprinted expression and its introns harbour no retrogenes. *Fam13c* is transcribed in oocytes and during early development. This could be important for the correct methylation of *Mcts2* in our knock-in construct, as transcription through the gene during embryonic development is thought to be important for the correct methylation of maternally imprinted genes. This has been shown at the imprinted *Gnas* [156] and *Snrpn* loci in mice [157]. Inserting *Mcts2* into *Fam13c* will allow us to test whether the use of alternative poly (A) sites is altered by the presence of *Mcts2*; that is, whether the introduction of *Mcts2* induces a different pattern of transcript expression relative to the endogenous *Fam13c* locus.

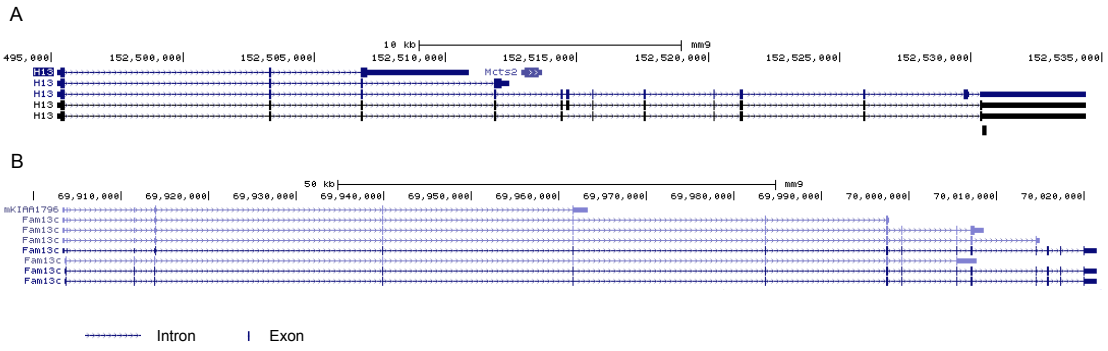


Figure 4.2 – Comparison of the *H13* and *Fam13c* gene structure. A – The gene structure of *H13* and *Mcts2*. B – The gene structure of *Fam13c*. Taken from the UCSC genome browser [155], using the mouse NCB137/mm9 reference genome (published in 2007.)

Both of these constructs contain homologous arms, which allow the vector to be targeted to the correct region of the genome (Figure 4.3). They also contain the gene for neomycin resistance and HSV thymidine kinase allowing for selection, and *lox P* sites allowing for Cre/*lox P* recombination [158]. Neomycin is used as a positive selection agent on the ES cells, as only the ES cells containing the construct are resistant to neomycin, while the ES cells that didn't integrate the construct are killed by the neomycin, leaving a population of ES cells that contain the construct. Gancyclovir

is used as a negative selection agent on the ES cells, as HSV thymidine kinase can convert gancyclovir to a toxic product killing the ES cells that have integrated the construct randomly rather than through homologous recombination at the target site. The *lox P* sites flank the promoter-associated CpG island of *Mcts2*, allowing it to be removed to prevent the expression of *Mcts2*. The Flp/FRT recombination system works in a similar way to the Cre/*lox P* system. The Flippase enzyme recognises the Flippase recognition target (FRT) sites, which in these constructs surround the neomycin resistance gene, allowing its removal [159], as this gene is not required in the mice.

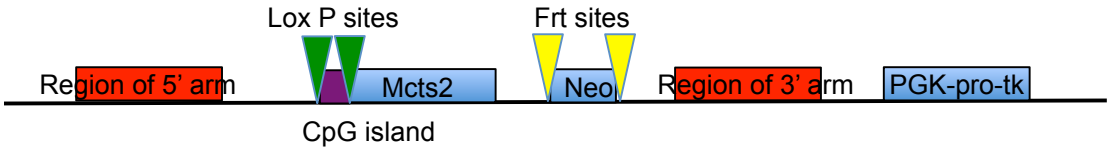


Figure 4.3 – Schematic of the knock-in/knock-out construct. For the knock-in construct the sequence in the region of the 5' and 3' arms is taken from the intron of *Fam13c* in which the construct is to be integrated. For the knock-out construct the sequence in the region of the 5' and 3' arms will be taken from the sequence of intron four of *H13* which surrounds *Mcts2*. *Lox P* sites allow for the removal of the CpG island of *Mcts2* when crossed with a Cre mouse. Frt sites allow for the removal of the neomycin resistance gene (labelled as neo) when crossed with a Flp mouse. PGK-pro-tk can convert gancyclovir to a toxic product, allowing for negative selection of cells that have integrated the construct randomly, rather than through homologous recombination at the target site.

Both constructs have been generated and electroporated into ES cells from a murine strain derived from 129. The cells generated were subjected to negative selection with HSV- thymidine kinase, which is toxic and due to its location in the construct would only be expressed in clones where insertion into the genome had been random rather than targeted. The surviving cells were subjected to positive selection through treatment with neomycin. Only the cells that had incorporated the correct part of the construct including the neomycin resistance gene would survive. The ES cells were generated by Xiangang Zou, a collaborator based at Cancer Research UK. My role in

the implementation of generation of *Mcts2* knock-in and knock-out mice was to check the ES cells for the correct integration of the two constructs, using Southern Blotting and long range PCR.

The second approach, the dose-responsive control of expression, uses two constructs based on the *H13/Mcts2* locus in a system that allows expression through these constructs to be controlled. Both of these constructs were designed and generated by myself. Due to the large size of *H13*, the DNA sequence utilised was limited to the first five exons (including the poly (A) sequence associated with exon five), and the first four introns. In place of *Mcts2* these constructs contain a reporter gene, mCherry, under the control of a tetracycline-responsive promoter. These constructs were transfected into tetracycline-responsive cell lines (also generated as part of this work). We hypothesised that varying transcription through mCherry should cause transcription through the rest of the construct to be altered. The constructs are described in greater detail in section 4.2.2. These experiments compliment those performed in the generation of *Mcts2* knock-in and knock-out mice.

Both of these approaches test the hypothesis that it is transcription from an internal promoter that causes premature polyadenylation of transcription through the host gene in a host/retrogene pair.

4.2. – Results

4.2.1. – Screening the ES Cells for Incorporation of the Constructs

The goals of generating *Mcts2* knock-in and knock-out mice were to demonstrate that intragenic promoters affect host gene transcript polyadenylation *in vitro* and *in vivo,* and

to demonstrate that relocating an intragenic promoter from its original position into an intron of a new host gene is sufficient to influence poly (A) site selection in the new host *in vitro*. To examine this property of intragenic promoters, we generated two constructs, one to knock-out *Mcts2* and one to knock-in *Mcts2* into another endogenous locus in the mouse genome. The new host gene was selected using the following criteria: the gene should be intronic, the intron structure should be similar to that of *H13*, and there should be no other intronic genes present in the new host gene. Based on these criteria *Fam13c* was chosen as the new host gene. The results presented here were generated by myself and reflect my contribution to this work. The design and generation of these constructs were undertaken by a post-doctoral fellow in the laboratory, Dr Mike Cowley.

Two Southern blots were performed to check for the insertion of the *Mcts2* knock-in construct into *Fam13c* via homologous recombination; one each to check that the 5' and 3' ends of the construct had been correctly integrated into the ES cell genome. To check that the 5' end of the construct had been correctly integrated into the ES cell genome DNA preparations from clones were digested with *Hind*III. The construct had been engineered to contain a *Hind*III site at its 5' end so that the fragment generated by digestion with this enzyme should be shorter than that obtained from the wild type *Fam13c* sequence. The digested DNA samples were electrophoresed through an agarose gel, followed by Southern blot transfer onto a nylon membrane and the fragments were detected using a radioactive probe that hybridises to a sequence present in both the construct and the mouse genome (Figure 4.4 A). Of the 14 clones screened, all apart from clone 96 had correctly integrated the 5' end of the construct, via homologous recombination, into the genome. The presence of double bands (Figure 4.4

B) shows that the construct had only been integrated into one of the two *Fam13c* alleles. To check that the 3' end of the construct had been correctly integrated into the ES cell genome, a second Southern blot was performed. For this test the clones were digested with *Stu*I, which due to the presence of another *Stu*I site in the construct should result in a larger fragment if the 3' end of the construct has been correctly integrated. A probe that hybridises to the neomycin resistance gene was used to detect the fragments (Figure 4.4 C). Of the 14 clones screened, all apart from clone 96 have correctly integrated the 3' end of the construct (Figure 4.4 D).

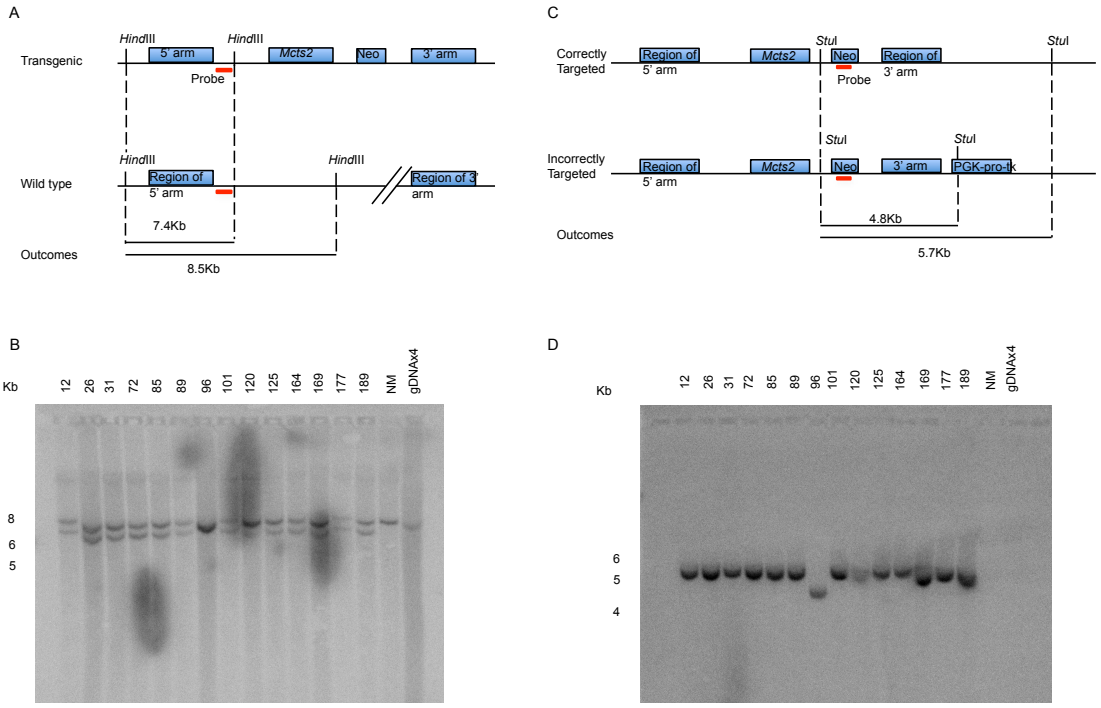


Figure 4.4 – Southern blot experiment to check for insertion of the *Mcts2* knock-in construct into *Fam13c*. A – Diagram to show the location of the *Hind*III restriction sites of interest and of the probe (red bar). B – Southern blot of clones digested with *Hind*III and hybridised with a probe to a sequence present both in the construct and in the mouse genome. C - Diagram to show the location of the *Stu*I restriction sites of interest and the probe used (red bar). D – Southern blot of clones digested with *Stu*I and incubated with a probe hybridising to the neomycin resistance gene. The numbers at the top of each column correlate to the clone it contains. gDNAx4 is a sample of genomic DNA loaded at 4 times the concentration of the test clones, acting as a control sample.

To check for the correct integration of the *Mcts2* knock-out construct into the ES cell genome, we used long-range PCR to screen the clones. Although theoretically we could use the same neomycin probe that had been used to check correct integration of the knock-in construct, we were unable to produce any Southern blot results using the probe targeted to the 5' end of the knock-out construct. I tried this several times with a range of probes specific for the 5' end of the construct, without any success. The reason for this remains unclear, although the successful hybridisation of other probes to different parts of the construct confirms that the DNA samples were of a good quality. We therefore decided to adopt an alternative approach based on PCR. Due to the large size of the construct two overlapping primer sets were used, both using a primer located in the genome close to the construct's integration site (Figure 4.5 A; see Appendix 7.2 for primer sequences). Of the six clones tested, we found just one clone (clone 272) had successfully integrated the *Mcts2* knock-out construct (Figure 4.5 B and C). This was verified using Sanger Sequencing.

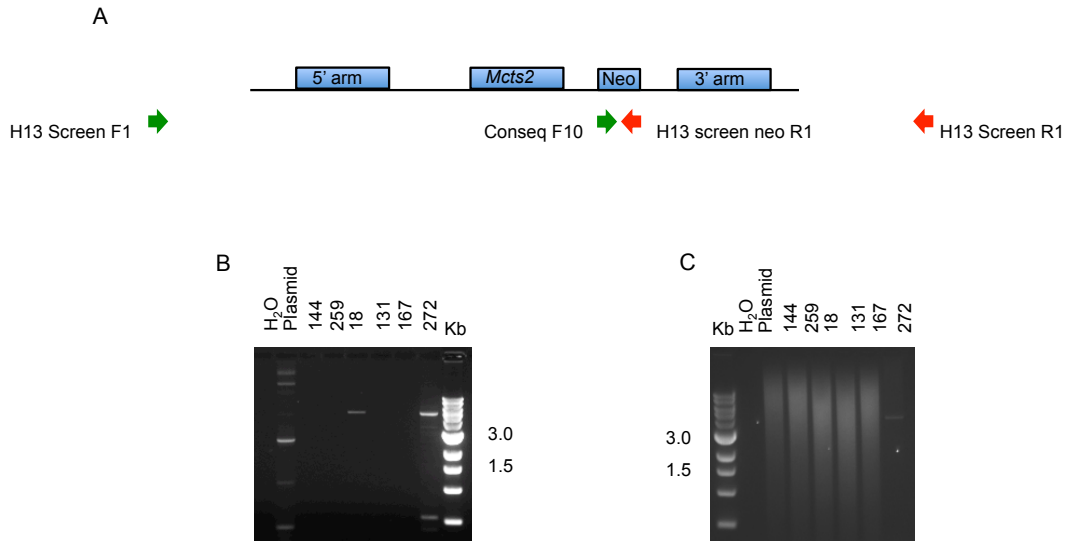


Figure 4.5 – Long range PCR to check for the insertion of the *Mcts2* knock-out construct into *H13*. A – Diagram of the *Mcts2* knock-out construct showing the location of the primers used for long range PCR. B – Gel of long range PCR using the H13 Screen F1 and H13 Screen neo R1 primers. The presence of a band at 4.9Kb in lanes containing clones 18 and 272 shows that these clones have incorporated the 5' end of the construct, from the sequence of the 5' arm to the neomycin resistance gene. C – Gel of long range PCR using the Conseq F10 and H13 Screen R1 primers. The presence of a band at 4.1Kb in the lane containing clone 272 shows that this is the only clone to have incorporated the 3' end of the construct, from the neomycin resistance gene to the sequence of the 3' arm. Primer sequences are in appendix 7.2.

Six of the clones which had successfully incorporated the *Mcts2* knock-in construct into *Fam13c* (clones 12, 31, 85, 89, 101, 164) and clone 272, which had successfully incorporated the *Mcts2* knock-out construct were chosen to be used in the generation of targeted transgenic mice. Several male animals containing each of these constructs were generated at the Transgenic Core Facility, CRUK Cambridge Research Institute. However, chimerism in these animals reached 65% as measured by coat colour, failing to reach the required minimum for founder mice at this facility of 70%. A second attempt to generate live animal models containing these constructs will be therefore be made after the completion of this thesis. However, we are able to work with the ES cells from the knock-in and knock-out lines screened by myself. This is currently the focus of work being performed by Samuele Amante (in Professor Oakey's laboratory);

Samuele is aiming to demonstrate that intragenic promoters affect host gene transcript polyadenylation in vitro. Together these experiments are expected to validate our existing correlative evidence by showing that intragenic promoters can control host gene polyadenylation, and that this occurs at endogenous loci in ES cells. This part of my thesis work therefore makes an important contribution to a larger ongoing team effort to understand these promoters.

4.2.2. – Design of a Construct to Test the Hypothesis that Alternative poly (A) Site Usage at the *H13/Mcts2* Locus is a Result of Transcriptional Interference

In the work described previously in this chapter, the constructs used allowed for the selection of ES cells through the addition of antibiotics to the medium. The antibiotic resistance genes the constructs contain are then removed through the crossing of the mice generated with Flp mice. However, as the approach we are using to test our hypothesis that alternative poly (A) site usage at the *H13*/*Mcts2* locus is a result of transcriptional interference will be performed in cell lines, the removal of these elements is not possible. We therefore took an alternative approach: we designed and engineered two expression vectors to modulate intragenic promoter activity and quantitatively measure the impacts on host gene polyadenylation.

We designed two versions of the construct (A and B), containing different complements of the *H13* exons and introns surrounding *Mcts2* (Figure 4.6). At the 5' end of both versions of the construct is a Ubiquitously acting Chromatin Opening Element (UCOE) [160], which is used to drive expression of the construct. It will also act to ensure that regardless of where the construct integrates in the genome it will be expressed, as UCOE acts to 'open up' the chromatin making it accessible to the DNA binding factors

needed for transcription. Downstream of this is the sequence of *H13* that surrounds *Mcts2* in its endogenous context. Both versions of the construct contain exons three and four, and introns three and four. Intron three has been included as this is retained in some *H13* transcripts. Intron four is the intron in which *Mcts2* is located, and it may contain sequences that are important for splicing or transcription of both *H13* and *Mcts2*. The portions of this intron which occur either side of the *Mcts2* sequence are therefore included in full. Exon five has also been included in both versions of the construct, and is fused to eGFP, allowing this to be used as a readout for expression from the H13 sequence. In addition to exons three to five of H13, construct A also contains exons one and two, while construct B does not. This makes construct B much smaller than construct A and we expected that this would make it easier to generate. In place of *Mcts2* both versions of the construct contain the sequence for the red fluorescent protein mCherry under the control of an inducible promoter, pTRE3G. pTRE3G consists of seven repeats of a 19bp tet operator sequence upstream of a Cytomegalovirus (CMV) promoter [161], and can be activated by the presence of tetracycline. This allows the expression of mCherry to be regulated through the addition of doxycycline (a tetracycline derivative) to the media of the growing cells.

Figure 4.6 – Diagram of the constructs for investigating the effect of an internal promoter on the transcription of the host gene. x1, x2, x3, x4 and x5 are the corresponding exons from *H13*. i3 and i4 are based on the corresponding introns in *H13*.

We designed the constructs for transfection into HEK 293 cells that are tetracycline-responsive. A synthetic tetracycline derivative, doxycycline, was used to activate the

pTRE3G promoter. Specifically, the addition of doxycycline to the cells causes a conformational change to the Tet-On 3G transactivator protein, allowing it to bind to the tet sequences in the pTRE3G promoter and activate expression of mCherry.

The constructs were designed such that transfection into tetracycline-responsive cell lines would allow us to determine the effect of transcription of the internal gene on poly (A) site usage of the host gene, through the expression of either mCherry or eGFP. If we are correct in our hypothesis that poly (A) site usage is fully determined by whether transcription of the internal gene occurs, we would expect to see eGFP expression when the cells are not exposed to tetracycline, because there should be no transcription from the pTRE3G promoter, allowing use of the downstream poly (A) sites; conversely in the presence of tetracycline, when the internal promoter is stimulated, we would expect to see expression of mCherry and a decrease in the expression of eGFP as some (if not all) transcripts terminate at a poly (A) site upstream of the minimal promoter in exon four (Figure 4.7).

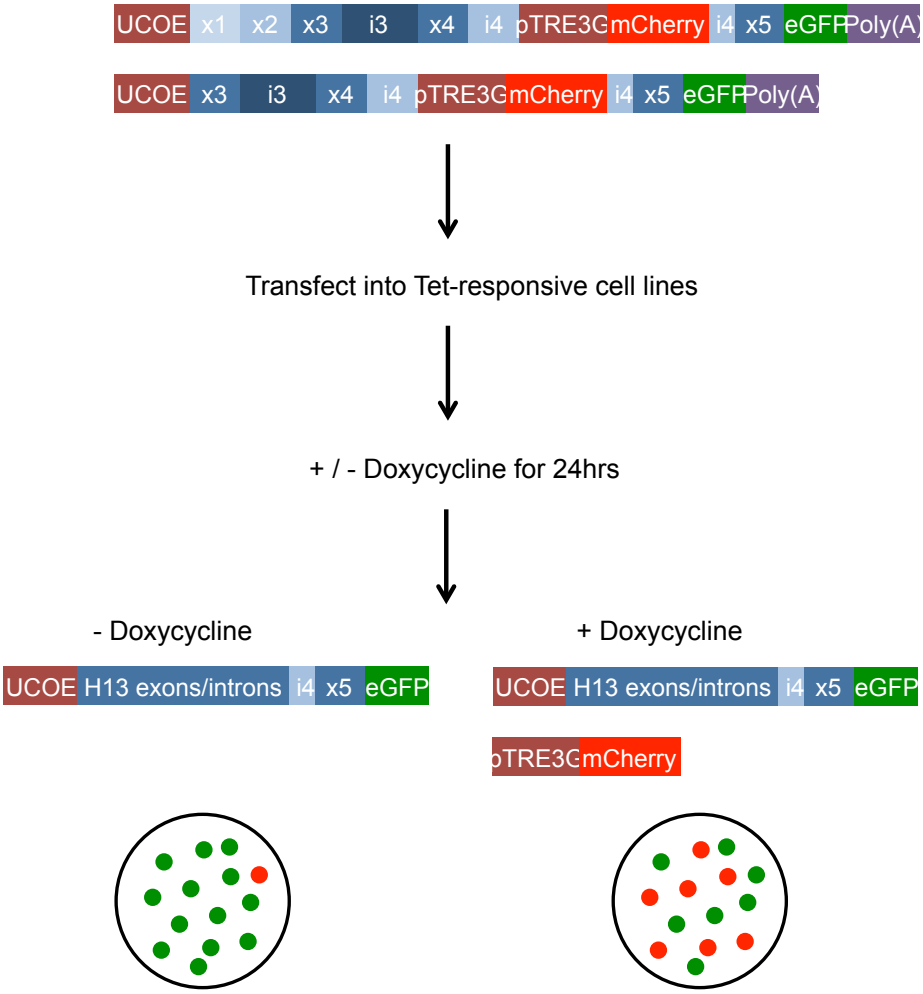


Figure 4.7 – A summary of expected outcomes when tet-responsive cell lines are treated with and without doxycycline for 24hrs. When cells are grown without doxycycline only the UCOE promoter should be active and so most transcripts should terminate at the poly (A) site at x5 (eGFP). When cells are grown in the presence of doxycycline for 24hrs both the UCOE and the pTRE3G promoters are active and so transcripts terminating with both the poly (A) site at x5 (eGFP) and mCherry are produced. x1, x2, x3, x4 and x5 are the corresponding exons from *H13*. i3 and i4 are based on the corresponding introns in *H13*.

4.2.3. – Generation of Tetracycline-responsive Cell Lines

Tetracycline-responsive cell lines were generated using the Tet-On 3G Inducible Expression System. HEK 293 cells were chosen because these are a human cell line: sequence differences between mouse and human therefore facilitate the differentiation

of endogenously expressed transcripts from those generated by our construct (which is derived from mouse sequence). The HEK 293 cells were transfected with pCMV-Tet3G and then selected for uptake with G418 (Life Technologies, cat number 11811023). To generate lines, the transfected cells were seeded onto plates at very low densities allowing individual colonies to form (taking anything from 1-4 weeks). These colonies were transferred into individual vessels and cultured, at 37^{o}C in 5% CO_2, until there were sufficient numbers (approximately 5 million cells) to test for expression of the Tet-On 3G transactivator and a high level of induction from pTRE3G.

To test the response of the cell lines to doxycycline, they were transfected with pTRE3G-Luc and TAL-Renilla, both with and without doxycycline. We then assayed the expression of firefly and *Renilla* luciferases using a Dual Luciferase Reporter Assay. Of the 12 HEK 293 lines tested only HEK 293-11 and HEK 293-14 showed a response to doxycycline (Figure 4.8). As doxycycline is added to the medium the expression of luciferase is stimulated in HEK 293-11 and HEK 293-14 compared to the parental HEK 293 line. This shows that these lines express the Tet-On 3G transactivator protein.

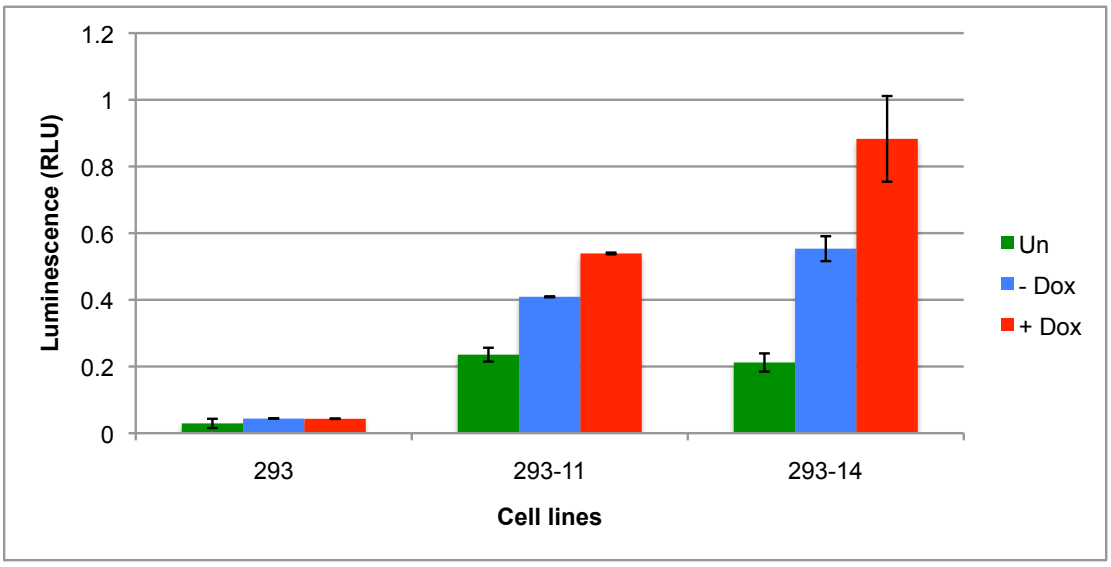


Figure 4.8 – The response of cell lines HEK 293, HEK 293-11 and HEK 293-14 to doxycycline. As doxycycline is added to the medium the expression of luciferase is stimulated in HEK 293-11 and HEK 293-14 compared to the parental HEK 293 line. This shows that these lines express the Tet-On 3G transactivator protein. Dox = doxycycline. n=3. The bars show +/- the standard error.

4.2.4. – Transfection of the Construct into Tetracycline-responsive Cell Lines

The two constructs were transiently transfected into the HEK 293-11 and HEK 293-14 lines, because these showed the best response to treatment with doxycycline (see section 2.4.11.3. in the Materials and Methods for more detail).

4.2.5. – Quantitative PCR to Assess Activity of Constructs

Quantitative PCR was used to assay expression of transcripts from the two constructs in both of the tetracycline-responsive cell lines generated, as shown in figure 4.9. We saw no difference in the expression levels of the transcripts generated from the constructs in either cell line in response to 24 hour treatment with doxycycline. We were expecting to see an increase in the transcripts detected by the mCherry assay and a decrease in the transcripts detected by the x5 and eGFP assays in the presence of doxycycline compared to in its absence. One possible explanation is that 24 hours of treatment is insufficient

to generate a difference in expression. Possible modifications to the procedure which we anticipate could produce the expected response include treatment with doxycycline over a longer time period [162] [163], or at a different concentration (for example at 24, 48, 72 and 96 hours at a concentration of 1, 2, 4, 6 and 10 µg/ml doxycycline); work to implement these modifications is ongoing in Professor Oakey's laboratory.

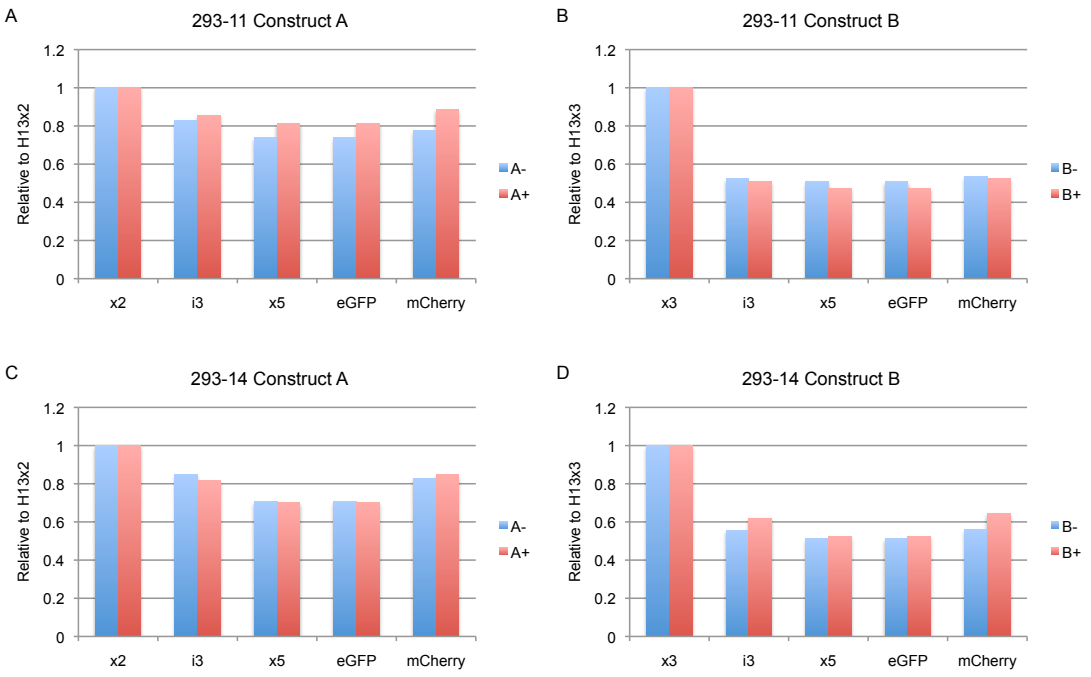


Figure 4.9 – qPCR showing expression of the different transcripts generated from constructs A and B. x2 is the region of H13 exon two, x3 is the region of H13 exon three, i3 is the region of H13 intron three and x5 is the region of H13 exon five in the constructs. x2 and x3 show the transcripts initiating from construct A and B respectively. i3 is in the transcripts that terminate upstream of mCherry. x5 and eGFP are in the transcripts that splice over mCherry. For A, B, C and D n=3.

Commercial fetal bovine serum added to the media used to generate and grow the cell lines can contain low levels of tetracycline, which can affect the results of these experiments. For this reason I used a tetracycline-free serum (available from Clontech) for the generation of these cell lines, and during the subsequent experiments.

Once the constructs were completed, and before they were used in any experiments, I used Sanger sequencing to determine the sequence of each construct. I compared the

sequences of the constructs generated to the planned sequences for each, and checked that each component was in the correct reading frame, and present in full. I also checked for mutations in the promoters and that the junctions between components were as planned. Therefore I feel confident to say that it is unlikely the unexpected results generated are due to a problem with the constructs themselves.

4.3. – Discussion

This chapter has described two complementary approaches designed to study imprinted gene expression at the *H13*/*Mcts2* locus: generation of *Mcts2* knock-in and knout-out mice, and the dose-responsive control of expression, in which two alternative constructs replicate the host/retrogene structure of *H13*/*Mcts2* and allow the expression of alternative transcripts of the host to be measured under controlled expression of the retrogene.

We have shown that good progress has been made towards the generation of *Mcts2* knock-in and knock-out mice. In section 4.2.1. I showed that several ES cell lines have successfully integrated the constructs, and work to culture these cells is ongoing in Professor Oakey's laboratory. These ES cells will be electroporated to allow for the uptake of flippase recombinase to remove the neomycin resistance gene from the construct. Experiments will be performed on these cells, as well as on cells that have been electroporated with *Cre* recombinase to remove the CpG island of *Mcts2*: comparison of these results should enable the effect of the CpG island on alternative poly (A) site usage on *H13* (in the knock-out model) or *Fam13c* (in the knock-in model) to be determined.

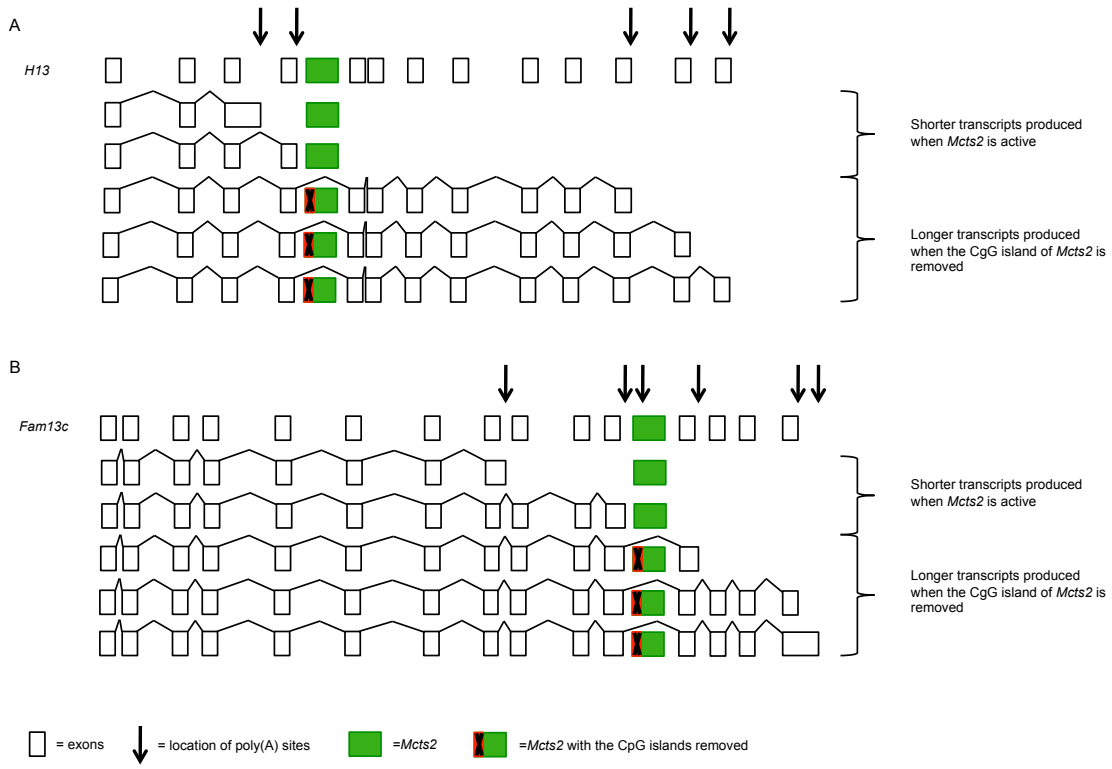


Figure 4.10 – Summary of the expected effect of active or inactive *Mcts2* on alternative poly (A) site usage in both *H13* and *Fam13c*. A – A diagram of the *H13* locus showing the location of *Mcts2* in intron four. Below this are the transcripts produced by using the different poly (A) sites present in the gene, in response to active or inactive *Mcts2*. B - A diagram of the *Fam13c* locus, showing the insertion of *Mcts2* in intron eleven. Below this are the transcripts produced by using the different poly (A) sites present in the gene, in response to active or inactive *Mcts2*.

For the knock-out construct (in which the endogenous *Mcts2* sequence is replaced by one where the CpG island is bound by *lox P* sites), if our hypothesis is correct we would expect to see an increase in the number of transcripts that terminate downstream of *Mcts2* when the CpG island of *Mcts2* is removed, inactivating the gene (see Figure 4.10 A). For the knock-in construct we are expecting that insertion of the active *Mcts2* will induce a preference for the use of the poly (A) sites upstream of the insertion site in *Fam13c*, and a reduction in the use of the downstream sites: when the CpG island of *Mcts2* is removed we are expecting the usage of poly (A) sites to reflect the usage seen in the wild-type locus, where *Mcts2* isn't present (Figure 4.10 B).

These experiments complement the dose-responsive control of expression, which uses two constructs in tet-responsive cell lines. Both approaches investigate the effect of an internal promoter on the expression of the gene in which it is located, but in slightly different ways. The ES cell experiments of mice experiments will generate results in a more biologically relevant system (as they are in a live animal model) whereas the dose-responsive control of expression experiments in tet-responsive cell lines allow us a more controlled environment in which to manipulate expression from the internal promoter and measure its influence on transcription.

I showed in section 2.4.10. of the Materials and Methods that we have successfully generated the construct, which should allow us to identify the effects of an internal promoter on the expression of the gene within which it is located. Two different versions of the construct were generated, differing in the region of *H13* (our host gene of interest) that they contain. The generation of these constructs was technically challenging, involving multiple cloning steps and the sourcing or generation of all the subunits. It was important to use unique restriction sites where possible to allow for the correct insertion (or removal if needed) of the individual subunits into the construct. Due to the number of cloning steps required this involved the use of some restriction enzymes requiring special conditions for their activity. We also successfully generated tetracycline-responsive cell lines, and demonstrated this by testing a subset for their response to doxycycline treatment (see section 4.2.3.). Of the twelve HEK 293 cell lines generated, only two showed a response to doxycycline. Unfortunately, the timescale of this thesis did not allow the next steps of this work to be taken, but there are several ways in which this work can be taken forward. Firstly, it could be

advantageous to test additional HEK 293 cell lines (and choose those that showed the best response to doxycycline, instead of using the only two that showed some response to doxycycline). It would also be beneficial to repeat the process of generating these cell lines again, and to generate an additional set of tetracycline-responsive cell lines, based on the 3T3 cell line.

The original goal of this part of the project was to generate cell lines that stably express our constructs in response to doxycycline treatment. Unfortunately, the numerous challenges that we encountered during the execution of the work prevented this goal from being fully achieved during the timescale of this thesis. However, I was able to perform some preliminary experiments in which the constructs were transiently transfected into the tetracycline-responsive cell lines that we did generate. Transiently transfected cells express the construct but it is not integrated into the genome, and so is not passed on during cell division. This means that transient transfections only express the construct for a short period of time (approximately 2-3 days), whereas in stable transfections the construct is expressed in the cells and is passed on to the progeny during cell division. Stable transfection therefore makes it possible to generate a cell line that stably expresses the construct, meaning that a single transfection is sufficient to produce a cell line that can be used for multiple experiments; conversely a transient transfection requires a separate transfection each time the experiment is repeated. Transient transfections are generally used when looking at short-term changes in gene expression or protein production. Stable transfections are useful for looking at longer-term changes, and at genetic regulation of the transfected DNA. Both systems will inform on our theory. The transient transfections will allow us to study expression through the construct and how this changes in response to doxycycline. When stable

transfections have been established, so that the construct is integrated into the genome, these should be informative regarding the mechanisms of gene regulation involved in alternative poly (A).

We observed no difference in expression across the construct in transfected cells treated with doxycycline (to activate expression through part of the construct) compared to untreated transfected cells (see section 4.2.5.). This was observed for both constructs in both of the tetracycline-responsive cells lines. There could be a number of reasons for this. Firstly, it is possible that additional work to optimise the transient transfection protocol would enable us to maximise expression from the constructs. It is possible that 24 hour's treatment with doxycycline is not long enough to have an effect on expression, or that the dose used needs to be higher. For example, human ES cells have been shown to respond to 2µg/ml of dox after 48 hours [162]. Alternatively, our experiments were designed with the aim of producing tetracycline-responsive cell lines that stably express the constructs: since the kit used to generate the cell lines was designed to generate stable cell lines, we may be unable to produce the results we expect with transient transfections. Therefore, these experiments need to be repeated in stable cell lines before any conclusions can be drawn.

There is previous work to support our hypothesis that the presence of an internal promoter can cause premature termination of its host gene [164]. A similar effect has been reported at another gene locus, where the imprinted *Nap1l5* causes premature termination of the *Herc3* gene in which it is located [143], which suggests that this mechanism could occur at other locations across the genome. Histone modifications and DNA methylation at intron/exon boundaries have also been shown to play a role in

alternative splicing [165], and it is possible that since we only used part of the *H13/Mcts2* locus these important marks were lost. It could be that the presence of an intragenic promoter alone is not sufficient to cause premature termination of the host gene.

Chapter 5

Discussion

5.1. - Overview

The work in this thesis aimed to investigate the mechanisms controlling gene expression, with particular emphasis on those acting on imprinted genes. We have focused on imprinted genes because even though the two alleles share an identical environment and are subject to the same influences, the gene is only expressed from one allele and not the other. Many imprinted genes are only expressed in an imprinted manner in certain tissues [166], and so the more we understand about the mechanisms regulating their gene expression the better we are able to understand tissue-specific regulation of genes. Two main approaches were used: one a genomewide approach using ChIP-Seq on a set of four DNA binding proteins, which have previously been shown to be associated with imprinted genes and their regulation, and a locus specific approach using a series of constructs to alter the expression of a model imprinted gene and its host, the *H13/Mcts2* locus. These two approaches were designed to be complimentary, with the ChIP-Seq investigating the interactions of these four proteins across the entire genome to look for trends, and the construct work using our well-studied model locus providing a mechanistic component to this research project, informing on the role of intragenic promoters in tissue-specific gene expression in mammals.

5.2. - Original Hypothesis

Chromosome looping plays an important role in the regulation of gene expression, by aiding the interactions between regulatory elements and gene promoters. Early

experiments at the *H19* imprinting control locus showed the importance of looping in *cis* regulation to control imprinted gene expression [167]. The development of more genomewide techniques detected novel *trans* interactions, widening the scope of known contacts with the *H19* DMD [168]. Both imprinted and non-imprinted gene interactions were detected, demonstrating that complex transcriptional networks were involved in the regulation of these genes and this has been further developed to reveal the co-regulation of imprinted genes in imprinted gene networks [169].

In the *H19* model these loops themselves serve to bring the regulatory elements into close enough proximity to the promoter of the gene so that they can affect transcription through it. Chromosome looping in general can occur in two main ways, through the interaction of two bound CTCF proteins with each other stabilised by the cohesin complex, and through interactions between cohesin and the Mediator Complex. Cohesin plays a key role in stabilising these chromosome loops. CTCF and cohesin are therefore intimately associated with the regulation of gene expression. ATRX and MeCP2 have been shown to co-localise with CTCF and cohesin at many loci across the genome, including at the *H19* ICR and the *Gtl2/Dlk1* imprinted regions [147]. As CTCF and cohesin ChIP-Seq datasets had already been generated in Professor Oakey's laboratory [22] it made sense to perform ChIP-Seq with ATRX and MeCP2 to investigate the interactions of these four proteins and their role in gene expression both genomewide and at all imprinted regions.

5.3. –Regulating Gene Expression

We have made some progress with elucidating the mechanisms responsible for regulating the expression of genes. In regards to imprinted genes it is clear from

published evidence that ATRX, MeCP2, CTCF and cohesin form a super complex over the DMR of the gene [170], but the details of how they interact to regulate gene expression once in place remains to be determined. Intronic promoters can influence alternative poly (A) site selection of their host gene, but more work needs to be done to determine whether their presence alone is enough to alter poly (A) site usage.

The ATRX and MeCP2 ChIP experiments required a great deal of optimisation to determine which antibodies, types of beads and samples worked together. It is clear to me now that I should have made the switch to using cell lines and ES cells much sooner than I did, but at the time I wanted to ensure a good match with the CTCF and cohesin ChIP-Seq datasets, which utilised murine brain. Switching to cell lines and ES cells sooner would have allowed me to repeat the CTCF and cohesin ChIP-Seq on these samples as well, ensuring that a match was maintained across the four datasets. It would have been beneficial to test a wider range of ATRX and MeCP2 antibodies, or to generate our own for the ChIP. The generation of the tet-responsive constructs was also time consuming, but went according to plan. The use of a new cloning kit, the Gibson Assembly Kit, was problematic, and in retrospect I spent too much time trying to optimise this approach, which ultimately resulted in the re-design of the cloning strategy to make use of more traditional cloning techniques, which worked well despite the large number of steps required. This extensive protocol development period left us with less time to optimise the expression of these constructs in cell lines in response to tetracycline than I would have liked.

5.4. – Complimentary and Related Approaches To Investigate The Regulation of Gene Expression

Active intragenic promoters influence alternative poly (A) site usage. *Mcts2* is a good model for studying this as it is an imprinted gene and so we have an active and an inactive promoter in the same cell type, and because the relationship between active *Mcts2* and alternative poly (A) site use by *H13* has already been established [146]. Several approaches are being used together to investigate the mechanisms of gene regulation at imprinted regions, including those described in this thesis investigating the influence of intragenic promoters on host poly (A) site usage, using *H13/Mcts2* locus as a model. We hypothesised that H3K36me3 is involved in poly (A) site selection. H3K36me3 affects the rate of transcription and one model for alternative poly (A) site selection that could be relevant is a competition based model, in which distal poly (A) sites are favoured, unless the relevant machinery can be bought to a proximal site before the polymerase reaches the distal site, in which case the proximal site will be used. By ablating the histone mark using siRNAs targeted to *Setd2*, which catalyses H3K36me3, we are altering histone methylation marks and can investigate if this mark is relevant to the mechanism of poly (A) site selection at the *H13/Mcts2* locus. To explore whether this is also relevant at other loci genomewide, and could indeed be a mechanism relevant to tissue-specific genes elsewhere, we have assayed the total transcriptome in ES cells using RNA-Seq. RNA-Seq allows the quantification of transcripts and the identification of structural elements, like intron-exon boundaries, as well as the ability to distinguish expression between specific alleles and alternate isoforms generated from the same gene [171]. By defining the expression pattern seen over the *H13/Mcts2* locus, we can search for this pattern of expression across the genome, allowing us to identify other regions where a gene located within an intron of a larger gene can alter its

transcription. Preliminary results from these data suggest that there are multiple sites across the genome that behave similarly and further interrogation of these data is ongoing in the laboratory.

5.5. - Future Work

In order to elucidate the mechanisms controlling imprinted gene expression, using the *H13*/*Mcts2* locus as a model, the work in this thesis needs to be extended, and considered with respect to the results of the other approaches being used by the Oakey laboratory. The ES cell work needs to be continued, and ideally *Mcts2* knock-in and knock-out mice need to be generated allowing for the correct setting of methylation marks through meiosis at the site of the construct, which could be important for regulation of transcription at this site. Stable tet-responsive cell lines need to be generated and the experiments described in chapter four repeated with the constructs that I generated. These will allow us to control expression through the constructs in a dose responsive manner. We expect these experiments to show that the presence of an internal promoter within a gene is enough to influence the use of alternative poly (A) sites on the host gene.

The ATRX ChIP-Seq data need to be analysed further in respect to ATRX binding allele-specifically at imprinted loci, and genomewide, to inform on its binding at these sites with CTCF and cohesin. This would inform on which of the two proposed models of interaction for these proteins (and MeCP2), if either, is more likely to be occurring, and if this alters depending on whether the gene is imprinted or not. The binding of MeCP2 at imprinted regions needs to be investigated, using an alternative method, since the ChIP-Seq approach has been unsuccessful. One option would to be to use data in

the public domain in related cell types. It would also be beneficial to look for the binding motifs of CTCF and cohesin within the peaks identified in the ATRX ChIP-Seq data, to see if they are binding at the same sites, or binding separately at the same regions.

Ultimately the ATRX ChIP needs to be optimised further to generate sequencing data of a high quality, which will aid its analysis, and allow us to form more confident conclusions on the binding of ATRX and its interactions with CTCF and cohesin. It is possible that the CTCF and cohesin ChIP-Seq experiments may need to be repeated to match the tissue type used for the ATRX ChIP-Seq. This will aid our comparisons of these datasets and ensure that we aren't missing interactions due to their transient nature in certain tissue types.

5.6. - Conclusions

The mechanisms controlling gene expression are complicated and dynamic. CTCF and cohesin have been shown to play a role in regulating gene expression through the establishment and maintenance of chromosome loops bringing regulatory regions into close proximity with gene promoters. ATRX and MeCP2 have been shown to co-localise with CTCF and cohesin at several sites across the genome. Due to the methylation specific binding of CTCF and MeCP2 (CTCF preferentially binds to un-methylated regions and MeCP2 binds to methylated regions) it is likely that these four proteins play a role in the regulation of imprinted gene expression. We have made a good start in investigating mechanisms controlling gene expression at imprinted regions, and those occurring more generally throughout the genome, but further work is needed to develop these.

Chapter 6

References

1. Bentley, D.L., *Coupling mRNA processing with transcription in time and space.* Nat Rev Genet, 2014. **15**(3): p. 163-75.
2. Mattick, J.S. and I.V. Makunin, *Small regulatory RNAs in mammals.* Hum Mol Genet, 2005. **14 Spec No 1**: p. R121-32.
3. Griffiths-Jones, S., et al., *miRBase: microRNA sequences, targets and gene nomenclature.* Nucleic Acids Res, 2006. **34**(Database issue): p. D140-4.
4. Kim, V.N., *MicroRNA biogenesis: coordinated cropping and dicing.* Nat Rev Mol Cell Biol, 2005. **6**(5): p. 376-85.
5. Mattick, J.S. and I.V. Makunin, *Non-coding RNA.* Hum Mol Genet, 2006. **15 Spec No 1**: p. R17-29.
6. Matera, A.G., R.M. Terns, and M.P. Terns, *Non-coding RNAs: lessons from the small nuclear and small nucleolar RNAs.* Nat Rev Mol Cell Biol, 2007. **8**(3): p. 209-20.
7. Orom, U.A., et al., *Long noncoding RNAs with enhancer-like function in human cells.* Cell, 2010. **143**(1): p. 46-58.
8. Yang, P.K. and M.I. Kuroda, *Noncoding RNAs and intranuclear positioning in monoallelic gene expression.* Cell, 2007. **128**(4): p. 777-86.
9. Costa, F.F., *Non-coding RNAs: Meet thy masters.* Bioessays, 2010. **32**(7): p. 599-608.
10. Palazzo, A.F. and E.S. Lee, *Non-coding RNA: what is functional and what is junk?* Front Genet, 2015. **6**: p. 2.
11. Lin, S., et al., *Nonallelic transcriptional roles of CTCF and cohesins at imprinted loci.* Mol Cell Biol, 2011. **31**(15): p. 3094-104.
12. McGrath, J. and D. Solter, *Completion of mouse embryogenesis requires both the maternal and paternal genomes.* Cell, 1984. **37**: p. 179-183.
13. Surani, M.A.H., S.C. Barton, and M.L. Norris, *Development of reconstituted mouse eggs suggests imprinting of the genome during gametogenesis.* Nature, 1984. **308**: p. 548-550.
14. Surani, M.A.H., S.C. Barton, and M.L. Norris, *Nuclear Transplantation in the Mouse: Heritable Differences between Parental Genomes after Activation of the Embryonic Genome.* Cell, 1986. **45**: p. 127-136.
15. Searle, A.G. and C.V. Beechey, *Genome imprinting phenomena on mouse chromosome 7.* Genetical Research, 1990. **56**(2-3): p. 237-244.
16. Cattanach, B.M., C.V. Beechey, and J. Peters, *Interactions between imprinting effects in the mouse.* Genetics, 2004. **168**(1): p. 397-413.
17. Ferguson-Smith, A.C., *Genomic imprinting: the emergence of an epigenetic paradigm.* Nat Rev Genet, 2011. **12**(8): p. 565-75.
18. Cattanach, B.M. and M. Kirk, *Differential activity of maternally and paternally derived chromosome regions in mice.* Nature, 1985. **315**: p. 496-498.

19. Cowley, M. and R.J. Oakey, *Resetting for the next generation.* Mol Cell, 2012. **48**(6): p. 819-21.
20. Williamson, C.M., et al. *World Wide Web Site - Mouse Imprinting Data and References - http://www.har.mrc.ac.uk/research/genomic_imprinting/*. 2013.
21. Morison, I.M., C.J. Paton, and S.D. Cleverley, *The imprinted gene and parent-of-origin effect database.* Nucleic Acids Research, 2001. **29**(1): p. 275-276.
22. Prickett, A.R., et al., *Genome-wide and parental allele-specific analysis of CTCF and cohesin DNA binding in mouse brain reveals a tissue-specific binding pattern and an association with imprinted differentially methylated regions.* Genome Res, 2013. **23**(10): p. 1624-35.
23. Thorvaldsen, J.L., K.L. Duran, and M.S. Bartolomei, *Deletion of the H19 differentially methylated domain results in loss of imprinted expression of H19 and Igf2.* Genes Dev, 1998. **12**: p. 3693-3702.
24. Nativio, R., et al., *Cohesin is required for higher-order chromatin conformation at the imprinted IGF2-H19 locus.* PLoS Genet, 2009. **5**(11): p. e1000739.
25. Fedoriw, A.M., et al., *Transgenic RNAi reveals essential function for CTCF in H19 gene imprinting.* Science, 2004. **303**: p. 238-240.
26. Kleinjan, D.A. and V. van Heyningen, *Long-Range Control of Gene Expression Emerging Mechanisms and Disruption in Disease.* Am J Hum Genet, 2005. **76**: p. 8-32.
27. Marsman, J. and J.A. Horsfield, *Long distance relationships: enhancer-promoter communication and dynamic gene transcription.* Biochim Biophys Acta, 2012. **1819**(11-12): p. 1217-27.
28. Merkenschlager, M. and D.T. Odom, *CTCF and cohesin: linking gene regulatory elements with their targets.* Cell, 2013. **152**(6): p. 1285-97.
29. Spielmann, M. and S. Mundlos, *Looking beyond the genes: the role of non-coding variants in human disease.* Hum Mol Genet, 2016.
30. Dixon, J.R., et al., *Topological domains in mammalian genomes identified by analysis of chromatin interactions.* Nature, 2012. **485**(7398): p. 376-80.
31. Lupianez, D.G., et al., *Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions.* Cell, 2015. **161**(5): p. 1012-25.
32. Illingworth, R.S. and A.P. Bird, *CpG islands--'a rough guide'.* FEBS Lett, 2009. **583**(11): p. 1713-20.
33. Gardiner-Garden, M. and M. Fromer, *CpG Islands in vertebrate genomes.* J Mol Biol, 1987. **196**: p. 261-282.
34. Blackledge, N.P. and R. Klose, *CpG island chromatin.* Epigenetics, 2011. **6**(2): p. 147-152.
35. Patil, V., R.L. Ward, and L.B. Hesson, *The evidence for functional non-CpG methylation in mammalian cells.* Epigenetics, 2014. **9**(6): p. 823-8.
36. Ziller, M.J., et al., *Genomic distribution and inter-sample variation of non-CpG methylation across human cell types.* PLoS Genet, 2011. **7**(12): p. e1002389.

37. Lister, R., et al., *Human DNA methylomes at base resolution show widespread epigenomic differences.* Nature, 2009. **462**(7271): p. 315-22.
38. Malone, C.S., et al., *CmC(A/T)GG DNA methylation in mature B cell lymphoma gene silencing.* Proc Natl Acad Sci U S A, 2001. **98**(18): p. 10404-9.
39. Inoue, S. and M. Oishi, *Effects of methylation of non-CpG sequence in the promoter region on the expression of human synaptotagmin XI (syt11).* Gene, 2005. **348**: p. 123-34.
40. Barres, R., et al., *Non-CpG methylation of the PGC-1alpha promoter through DNMT3B controls mitochondrial density.* Cell Metab, 2009. **10**(3): p. 189-98.
41. Xie, W., et al., *Base-resolution analyses of sequence and parent-of-origin dependent DNA methylation in the mouse genome.* Cell, 2012. **148**(4): p. 816-31.
42. Moore, L.D., T. Le, and G. Fan, *DNA methylation and its basic function.* Neuropsychopharmacology, 2013. **38**(1): p. 23-38.
43. Bourc'his, D., et al., *Dnmt3L and the establishment of maternal genomic imprints.* Science, 2001. **294**(5551): p. 2536-9.
44. Elliott, E.N., et al., *Dnmt1 is essential to maintain progenitors in the perinatal intestinal epithelium.* Development, 2015. **142**(12): p. 2163-72.
45. Uysal, F., G. Akkoyunlu, and S. Ozturk, *Dynamic expression of DNA methyltransferases (DNMTs) in oocytes and early embryos.* Biochimie, 2015. **116**: p. 103-13.
46. Okano, M., et al., *DNA Methyltransferases Dnmt3a and Dnmt3b Are Essential for De Novo Methylation*
and Mammalian Development. Cell, 1999. **99**: p. 247-257.
47. Bourc'his, D. and T.H. Bestor, *Meiotic catastrophe and retrotransposon reactivation in male germ cells lacking Dnmt3L.* Nature, 2004. **431**: p. 96-99.
48. He, Y.F., et al., *Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA.* Science, 2011. **333**(6047): p. 1303-7.
49. Ito, S., et al., *Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine.* Science, 2011. **333**(6047): p. 1300-3.
50. Reik, W., W. Dean, and J. Walter, *Epigenetic reprogramming in mammalian development.* Science, 2001. **293**(5532): p. 1089-93.
51. Lucifero, D., et al., *Coordinate regulation of DNA methyltransferase expression during oogenesis.* BMC Dev Biol, 2007. **7**: p. 36.
52. Pfeifer, G.P., Kadam, S., Jin, S-G., *5-hydroxymethylcytosine and its potential roles in development and cancer.* Epigenetics and Chromatin, 2013. **6**(10).
53. Song, C.X. and C. He, *Potential functional roles of DNA demethylation intermediates.* Trends Biochem Sci, 2013. **38**(10): p. 480-4.
54. Plongthongkum, N., D.H. Diep, and K. Zhang, *Advances in the profiling of DNA modifications: cytosine methylation and beyond.* Nat Rev Genet, 2014. **15**(10): p. 647-61.

55. Sun, Z., et al., *High-resolution enzymatic mapping of genomic 5-hydroxymethylcytosine in mouse embryonic stem cells.* Cell Rep, 2013. **3**(2): p. 567-76.
56. Booth, M.J., E.A. Raiber, and S. Balasubramanian, *Chemical methods for decoding cytosine modifications in DNA.* Chem Rev, 2015. **115**(6): p. 2240-54.
57. Frommer, M., McDonald, L. E., Millar, D. S., Collis, C. M., Watt, F., Grigg, G. W., Molloy, P. L., Paul, C. L., *A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands.* Proc Natl Acad Sci U S A, 1992. **89**: p. 1827- 1831.
58. Delaney, C., S.K. Garg, and R. Yung, *Analysis of DNA Methylation by Pyrosequencing.* Methods Mol Biol, 2015. **1343**: p. 249-64.
59. Ulahannan, N., Greally, J. M., *Genome-wide assays that identify and quantify modified cytosines in human disease studies.* Epigenetics and Chromatin, 2015. **8**(5).
60. Yu, M., et al., *Base-resolution analysis of 5-hydroxymethylcytosine in the mammalian genome.* Cell, 2012. **149**(6): p. 1368-80.
61. Booth, M.J., et al., *Oxidative bisulfite sequencing of 5-methylcytosine and 5-hydroxymethylcytosine.* Nat Protoc, 2013. **8**(10): p. 1841-51.
62. Lu, X., et al., *Chemical modification-assisted bisulfite sequencing (CAB-Seq) for 5-carboxylcytosine detection in DNA.* J Am Chem Soc, 2013. **135**(25): p. 9315-7.
63. Booth, M.J., et al., *Quantitative sequencing of 5-formylcytosine in DNA at single-base resolution.* Nat Chem, 2014. **6**(5): p. 435-40.
64. Allen, B.L. and D.J. Taatjes, *The Mediator complex: a central integrator of transcription.* Nat Rev Mol Cell Biol, 2015. **16**(3): p. 155-66.
65. Carlsten, J.O., X. Zhu, and C.M. Gustafsson, *The multitalented Mediator complex.* Trends Biochem Sci, 2013. **38**(11): p. 531-7.
66. Richard, P. and J.L. Manley, *Transcription termination by nuclear RNA polymerases.* Genes Dev, 2009. **23**(11): p. 1247-69.
67. Tammen, S.A., S. Friso, and S.W. Choi, *Epigenetics: the link between nature and nurture.* Mol Aspects Med, 2013. **34**(4): p. 753-64.
68. Fatemi, M. and P.A. Wade, *MBD family proteins: reading the epigenetic code.* J Cell Sci, 2006. **119**(Pt 15): p. 3033-7.
69. Baubec, T., et al., *Methylation-dependent and -independent genomic targeting principles of the MBD protein family.* Cell, 2013. **153**(2): p. 480-92.
70. Du, Q., et al., *Methyl-CpG-binding domain proteins readers of the epigenome.* Epigenomics, 2015. **7**(6): p. 1051-1073.
71. Nan, X., et al., *Interaction between chromatin proteins MECP2 and ATRX is disrupted by mutations that cause inherited mental retardation.* Proc Natl Acad Sci U S A, 2007. **104**(8): p. 2709-14.
72. Gigek, C.O., E.S. Chen, and M.A. Smith, *Methyl-CpG-Binding Protein (MBD) Family: Epigenomic Read-Outs Functions and Roles in Tumorigenesis and Psychiatric Diseases.* J Cell Biochem, 2016. **117**(1): p. 29-38.

73. Long, H.K., N.P. Blackledge, and R.J. Klose, *ZF-CxxC domain-containing proteins, CpG islands and the chromatin connection.* Biochem Soc Trans, 2013. **41**(3): p. 727-40.
74. Blackledge, N.P., et al., *CpG Islands Recruit a Histone H3 Lysine 36 Demethylase.* Mol Cell, 2010. **38**(2-2): p. 179-190.
75. Blackledge, N.P., J.P. Thomson, and P.J. Skene, *CpG island chromatin is shaped by recruitment of ZF-CxxC proteins.* Cold Spring Harb Perspect Biol, 2013. **5**(11): p. a018648.
76. Ooi, S.K., et al., *DNMT3L connects unmethylated lysine 4 of histone H3 to de novo methylation of DNA.* Nature, 2007. **448**(7154): p. 714-7.
77. Thomson, J.P., et al., *CpG islands influence chromatin structure via the CpG-binding protein Cfp1.* Nature, 2010. **464**(7291): p. 1082-6.
78. Kim, S., N.K. Yu, and B.K. Kaang, *CTCF as a multifunctional protein in genome regulation and gene expression.* Exp Mol Med, 2015. **47**: p. e166.
79. Fang, R., et al., *Functional diversity of CTCFs is encoded in their binding motifs.* BMC Genomics, 2015. **16**: p. 649.
80. Holwerda, S.J. and W. de Laat, *CTCF: the protein, the binding partners, the binding sites and their chromatin loops.* Philos Trans R Soc Lond B Biol Sci, 2013. **368**(1620): p. 20120369.
81. Kim, A. and A. Dean, *Chromatin loop formation in the beta-globin locus and its role in globin gene transcription.* Mol Cells, 2012. **34**(1): p. 1-5.
82. Splinter, E., et al., *CTCF mediates long-range chromatin looping and local histone modification in the beta-globin locus.* Genes Dev, 2006. **20**(17): p. 2349-54.
83. Singh, V.P. and J.L. Gerton, *Cohesin and human disease: lessons from mouse models.* Curr Opin Cell Biol, 2015. **37**: p. 9-17.
84. Xiao, T., J. Wallace, and G. Felsenfeld, *Specific sites in the C terminus of CTCF interact with the SA2 subunit of the cohesin complex and are required for cohesin-dependent insulation activity.* Mol Cell Biol, 2011. **31**(11): p. 2174-83.
85. Watrin, E. and J.M. Peters, *Cohesin and DNA damage repair.* Exp Cell Res, 2006. **312**(14): p. 2687-93.
86. Dorsett, D. and M. Merkenschlager, *Cohesin at active genes: a unifying theme for cohesin and gene expression from model organisms to humans.* Curr Opin Cell Biol, 2013. **25**(3): p. 327-33.
87. Parelho, V., et al., *Cohesins functionally associate with CTCF on mammalian chromosome arms.* Cell, 2008. **132**(3): p. 422-33.
88. Wendt, K.S., et al., *Cohesin mediates transcriptional insulation by CCCTC-binding factor.* Nature, 2008. **451**(7180): p. 796-801.
89. Mehta, G.D., et al., *Cohesin: functions beyond sister chromatid cohesion.* FEBS Lett, 2013. **587**(15): p. 2299-312.
90. Remeseiro, S., A. Cuadrado, and A. Losada, *Cohesin in development and disease.* Development, 2013. **140**(18): p. 3715-8.
91. Skibbens, R.V., et al., *Cohesinopathies of a feather flock together.* PLoS Genet, 2013. **9**(12): p. e1004036.
92. Gibbons, R.J., et al., *Mutations in the chromatin-associated protein ATRX.* Hum Mutat, 2008. **29**(6): p. 796-802.

93. Medina, C.F., et al., *Altered visual function and interneuron survival in Atrx knockout mice: inference for the human syndrome.* Hum Mol Genet, 2009. **18**(5): p. 966-77.
94. Clynes, D., et al., *ATRX dysfunction induces replication defects in primary mouse cells.* PLoS One, 2014. **9**(3): p. e92915.
95. Voon, H.P., et al., *ATRX Plays a Key Role in Maintaining Silencing at Interstitial Heterochromatic Loci and Imprinted Genes.* Cell Rep, 2015. **11**(3): p. 405-18.
96. Bérubé, N.G., C.A. Smeenk, and D.J. Picketts, *Cell cycle-dependent phosphorylation of the ATRX protein correlates with changes in nuclear matrix and chromatin association.* Hum Mol Genet, 2000. **9**(4): p. 539-547.
97. Clynes, D., D.R. Higgs, and R.J. Gibbons, *The chromatin remodeller ATRX: a repeat offender in human disease.* Trends Biochem Sci, 2013. **38**(9): p. 461-6.
98. Klose, R.J., et al., *DNA binding selectivity of MeCP2 due to a requirement for A/T sequences adjacent to methyl-CpG.* Mol Cell, 2005. **19**(5): p. 667-78.
99. Long, S.W., et al., *A brain-derived MeCP2 complex supports a role for MeCP2 in RNA processing.* Biosci Rep, 2011. **31**(5): p. 333-43.
100. Lyst, M.J. and A. Bird, *Rett syndrome: a complex disorder with simple roots.* Nat Rev Genet, 2015. **16**(5): p. 261-75.
101. Skene, P.J., et al., *Neuronal MeCP2 is expressed at near histone-octamer levels and globally alters the chromatin state.* Mol Cell, 2010. **37**(4): p. 457-68.
102. Cuddapah, V.A., et al., *Methyl-CpG-binding protein 2 (MECP2) mutation type is associated with disease severity in Rett syndrome.* J Med Genet, 2014. **51**(3): p. 152-8.
103. Christodoulou, J., et al., *RettBASE: The IRSA MECP2 Variation Database - A New Mutation Database in Evolution.* Hum Mutat, 2003. **21**: p. 466-472.
104. Cheema, M.S. and J. Ausio, *The Structural Determinants behind the Epigenetic Role of Histone Variants.* Genes (Basel), 2015. **6**(3): p. 685-713.
105. Marzluff, W.F., et al., *The Human and Mouse Replication-Dependent Histone Genes.* Genomics, 2002. **80**(5): p. 487-498.
106. Harshman, S.W., et al., *H1 histones: current perspectives and challenges.* Nucleic Acids Res, 2013. **41**(21): p. 9593-609.
107. Khare, S.P., et al., *HIstome: a relational knowledgebase of human histone proteins and histone modifying enzymes.* Nucleic Acids Research, 2011(Database issue doi:10.1093/nar/gkr1125).
108. Sarma, K. and D. Reinberg, *Histone variants meet their match.* Nat Rev Mol Cell Biol, 2005. **6**(2): p. 139-49.
109. Draizen, E.J., et al., *HistoneDB 2.0: a histone database with variants--an integrated resource to explore histones and their variants.* Database (Oxford), 2016. **2016**.
110. Santoro, S.W. and C. Dulac, *Histone variants and cellular plasticity.* Trends Genet, 2015. **31**(9): p. 516-27.

111. Molden, R.C., et al., *Multi-faceted quantitative proteomics analysis of histone H2B isoforms and their modifications.* Epigenetics Chromatin, 2015. **8**: p. 15.
112. Goldberg, A.D., et al., *Distinct factors control histone variant H3.3 localization at specific genomic regions.* Cell, 2010. **140**(5): p. 678-91.
113. Kamakaka, R.T., Biggins, S., *Histone variants: deviants?* Genome Res, 2005. **19**(3): p. 295-310.
114. Strahl, B.D. and C.D. Allis, *The language of covalent histone modifications.* Nature, 2000. **403**: p. 41-45.
115. Bannister, A.J. and T. Kouzarides, *Regulation of chromatin by histone modifications.* Cell Res, 2011. **21**(3): p. 381-95.
116. Gardner, K.E., C.D. Allis, and B.D. Strahl, *Operating on chromatin, a colorful language where context matters.* J Mol Biol, 2011. **409**(1): p. 36-46.
117. Taverna, S.D., et al., *How chromatin-binding modules interpret histone modifications: lessons from professional pocket pickers.* Nat Struct Mol Biol, 2007. **14**(11): p. 1025-40.
118. Campos, E.I. and D. Reinberg, *Histones: annotating chromatin.* Annu Rev Genet, 2009. **43**: p. 559-99.
119. Zhou, V.W., A. Goren, and B.E. Bernstein, *Charting histone modifications and the functional organization of mammalian genomes.* Nat Rev Genet, 2011. **12**(1): p. 7-18.
120. Bloom, K.S., *Centromeric heterochromatin: the primordial segregation machine.* Annu Rev Genet, 2014. **48**: p. 457-84.
121. Li, G., et al., *Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation.* Cell, 2012. **148**(1-2): p. 84-98.
122. Doyle, B., et al., *Chromatin loops as allosteric modulators of enhancer-promoter interactions.* PLoS Comput Biol, 2014. **10**(10): p. e1003867.
123. Aragon, L., E. Martinez-Perez, and M. Merkenschlager, *Condensin, cohesin and the control of chromatin states.* Curr Opin Genet Dev, 2013. **23**(2): p. 204-11.
124. Zuin, J., et al., *Cohesin and CTCF differentially affect chromatin architecture and gene expression in human cells.* Proc Natl Acad Sci U S A, 2014. **111**(3): p. 996-1001.
125. Poss, Z.C., C.C. Ebmeier, and D.J. Taatjes, *The Mediator complex and transcription regulation.* Crit Rev Biochem Mol Biol, 2013. **48**(6): p. 575-608.
126. Bonora, G., K. Plath, and M. Denholtz, *A mechanistic link between gene regulation and genome architecture in mammalian development.* Curr Opin Genet Dev, 2014. **27**: p. 92-101.
127. Rao, S.S., et al., *A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping.* Cell, 2014. **159**(7): p. 1665-80.
128. Sexton, T. and G. Cavalli, *The role of chromosome domains in shaping the functional genome.* Cell, 2015. **160**(6): p. 1049-59.
129. Plasschaert, R.N. and M.S. Bartolomei, *Tissue-specific regulation and function of Grb10 during growth and neuronal commitment.* Proc Natl Acad Sci U S A, 2015. **112**(22): p. 6841-6847.

130. Pan, Q., et al., *Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing.* Nat Genet, 2008. **40**(12): p. 1413-5.
131. Lev Maor, G., A. Yearim, and G. Ast, *The alternative role of DNA methylation in splicing regulation.* Trends Genet, 2015. **31**(5): p. 274-80.
132. Ast, G., *How did alternative splicing evolve?* Nat Rev Genet, 2004. **5**(10): p. 773-82.
133. Luco, R.F., et al., *Epigenetics in alternative pre-mRNA splicing.* Cell, 2011. **144**(1): p. 16-26.
134. Shukla, S., et al., *CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing.* Nature, 2011. **479**(7371): p. 74-9.
135. Maunakea, A.K., et al., *Intragenic DNA methylation modulates alternative splicing by recruiting MeCP2 to promote exon recognition.* Cell Res, 2013. **23**(11): p. 1256-69.
136. Proudfoot, N.J., *Ending the message: poly(A) signals then and now.* Genes Dev, 2011. **25**(17): p. 1770-82.
137. Tian, B., et al., *A large-scale analysis of mRNA polyadenylation of human and mouse genes.* Nucleic Acids Res, 2005. **33**(1): p. 201-12.
138. Kaer, K., et al., *Intronic L1 retrotransposons and nested genes cause transcriptional interference by inducing intron retention, exonization and cryptic polyadenylation.* PLoS One, 2011. **6**(10): p. e26099.
139. Shearwin, K.E., B.P. Callen, and J.B. Egan, *Transcriptional interference--a crash course.* Trends Genet, 2005. **21**(6): p. 339-45.
140. Lis, M. and D. Walther, *The orientation of transcription factor binding site motifs in gene promoter regions: does it matter?* BMC Genomics, 2016. **17**: p. 185.
141. Wei, W., et al., *Functional consequences of bidirectional promoters.* Trends Genet, 2011. **27**(7): p. 267-76.
142. McCole, R.B. and R.J. Oakey, *Unwitting hosts fall victim to imprinting.* Epigenetics, 2008. **3**(5): p. 258-260.
143. Cowley, M., et al., *Epigenetic control of alternative mRNA processing at the imprinted Herc3/Nap1l5 locus.* Nucleic Acids Res, 2012. **40**(18): p. 8917-26.
144. Fuks, F., *The DNA methyltransferases associate with HP1 and the SUV39H1 histone methyltransferase.* Nucleic Acids Research, 2003. **31**(9): p. 2305-2312.
145. Williamson, C.M., et al., *Identification of an imprinting control region affecting the expression of all transcripts in the Gnas cluster.* Nat Genet, 2006. **38**(3): p. 350-5.
146. Wood, A.J., et al., *Regulation of alternative polyadenylation by genomic imprinting.* Genes Dev, 2008. **22**(9): p. 1141-6.
147. Kernohan, K.D., et al., *ATRX partners with cohesin and MeCP2 and contributes to developmental silencing of imprinted genes in the brain.* Dev Cell, 2010. **18**(2): p. 191-202.
148. Langmead, B. and S.L. Salzberg, *Fast gapped-read alignment with Bowtie 2.* Nat Methods, 2012. **9**(4): p. 357-9.

149. Langmead, B., et al., *Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.* Genome Biol, 2009. **10**(3): p. R25.
150. Rubio, E.D., et al., *CTCF physically links cohesin to chromatin.* Proc Natl Acad Sci U S A, 2008. **105**(24): p. 8309-14.
151. McCole, R.B., *Evolution and Regulation of Imprinted Retrogenes.* PhD thesis, 2010.
152. Maupetit-Mehouas, S., et al., *Imprinting control regions (ICRs) are marked by mono-allelic bivalent chromatin when transcriptionally inactive.* Nucleic Acids Res, 2015.
153. *http://www.phrap.com/phred/.*
154. Law, M.J., et al., *ATR-X syndrome protein targets tandem repeats and influences allele-specific expression in a size-dependent manner.* Cell, 2010. **143**(3): p. 367-78.
155. Kent, W.J., et al., *The human genome browserat UCSC.* Genome Res, 2002. **12**(6): p. 996-1006.
156. Chotalia, M., et al., *Transcription is required for establishment of germline methylation marks at imprinted genes.* Genes Dev, 2009. **23**(1): p. 105-17.
157. Smith, E.Y., et al., *Transcription is required to establish maternal imprinting at the Prader-Willi syndrome and Angelman syndrome locus.* PLoS Genet, 2011. **7**(12): p. e1002422.
158. Sauer, B., *Inducible Gene Targeting in Mice Using the Cre/lox system.* Methods, 1998. **14**(4): p. 381-92.
159. Bouabe, H. and K. Okkenhaug, *Gene targeting in mice: a review.* Methods Mol Biol, 2013. **1064**: p. 315-36.
160. Zhang, F., et al., *Lentiviral vectors containing an enhancer-less ubiquitously acting chromatin opening element (UCOE) provide highly reproducible and stable transgene expression in hematopoietic cells.* Blood, 2007. **110**(5): p. 1448-57.
161. Clontech Laboratories, I., *Tet-On 3G Inducible Expression Systems User Manual.*
162. Ozair, M.Z., et al., *SMAD7 directly converts human embryonic stem cells to telencephalic fate by a default mechanism.* Stem Cells, 2013. **31**(1): p. 35-47.
163. Villasenor, A., et al., *EphB3 marks delaminating endocrine progenitor cells in the developing pancreas.* Dev Dyn, 2012. **241**(5): p. 1008-19.
164. Monk, D., et al., *Human imprinted retrogenes exhibit non-canonical imprint chromatin signatures and reside in non-imprinted host genes.* Nucleic Acids Res, 2011. **39**(11): p. 4577-86.
165. Podlaha, O., et al., *Histone modifications are associated with transcript isoform diversity in normal and cancer cells.* PLoS Comput Biol, 2014. **10**(6): p. e1003611.
166. Prickett, A.R. and R.J. Oakey, *A survey of tissue-specific genomic imprinting in mammals.* Mol Genet Genomics, 2012. **287**(8): p. 621-30.
167. Murrell, A., S. Heeson, and W. Reik, *Interaction between differentially methylated regions partitions the imprinted genes Igf2 and H19 into parent-specific chromatin loops.* Nat Genet, 2004. **36**(8): p. 889-93.

168. Zhao, Z., et al., *Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions.* Nat Genet, 2006. **38**(11): p. 1341-7.
169. Varrault, A., et al., *Zac1 regulates an imprinted gene network critically involved in the control of embryonic growth.* Dev Cell, 2006. **11**(5): p. 711-22.
170. Kernohan, K.D., et al., *Analysis of neonatal brain lacking ATRX or MeCP2 reveals changes in nucleosome density, CTCF binding and chromatin looping.* Nucleic Acids Res, 2014. **42**(13): p. 8356-68.
171. Wang, Z., M. Gerstein, and M. Snyder, *RNA-Seq: a revolutionary tool for transcriptomics.* Nat Rev Genet, 2009. **10**(1): p. 57-63.

Chapter 7

Appendix

7.1. - Buffers and Reagents

2.3.2. - Chromatin Immunoprecipitation – Method 1

Dilution Buffer (250ml)	Volume (ml)
1M Tris-HCl pH8	4.175
5M NaCl	8.250
20% Triton X-100	13.750
0.5M EDTA	0.600
H_2O	223.125

Wash Buffer 1 (500ml)	Volume (ml) (final concentration)
1M Tris-HCl pH8	10.0 (20mM)
5M NaCl	15.0 (150mM)
10% SDS	5.0 (0.1%)
20% Triton X-100	25.0 (1%)
0.5M EDTA	2.0 (2mM)
0.1M PMSF (in 100% Ethanol)	0.2
H_2O	442.8

Wash Buffer 2 (500ml)	Volume (ml) (final concentration)
1M Tris-HCl pH8	10.0 (20mM)
5M NaCl	50.0 (500mM)
10% SDS	5.0 (0.1%)
20% Triton X-100	25.0 (1%)
0.5M EDTA	2.0 (2mM)
0.1M PMSF (in 100% Ethanol)	0.2
H_2O	407.8

Wash Buffer 3 (500ml)	Volume (ml) (final concentration)
1M Tris-HCl pH8	5.0 (10mM)
1M lithium chloride	125.0 (250mM)
20% IpeGal 360	25.0 (1%)
20% sodium deoxycholate	25.0 (2g)
0.5M EDTA	1.0 (1mM)
0.1M PMSF (in 100% Ethanol)	0.2
H_2O	318.8

EDTA-free protease inhibitor (Roche, cat number 04693132001) was added to the wash buffers just before use.

2.3.4. - Chromatin Immunoprecipitation – Method 3

PBS/BSA (50ml)	Volume (ml) (final concentration)
BSA	0.25g (5mg/ml)
PBS	50.00

This should be made up fresh each time and kept cold.

Lysis Buffer 1 (100ml)	Volume (ml) (final concentration)
1M HEPES-KOH, pKa 7.5	10.00 (100mM)
5M NaCl	2.80 (140mM)
0.5M EDTA	0.20 (1mM)
Glycerol	10.00 (10%)
100% NP-40	0.50 (0.5%)
Triton X-100	0.25 (0.25%)

Lysis Buffer 2 (100ml)	Volume (ml) (final concentration)
5M NaCl	4.0 (200mM)
0.5M EDTA	0.2 (1mM)
125mM EGTA	0.4 (0.5mM)
500mM Tris pH8	2.0 (1mM)

Lysis Buffer 3 (100ml)	Volume (ml) (final concentration)
0.5M EDTA	0.2 (1mM)
125mM EGTA	0.4 (0.5mM)
500mM Tris pH8	2.0 (1mM)
5M NaCl	2.0 (100mM)
Na-Deoxycholate	0.1g (0.1%)
N-lauroyl sarcosine	0.5g (0.5%)

Wash Buffer (RIPA) (100ml)	Volume (ml) (final concentration)
500mM HEPES, pKa 7.4	10.00 (50mM, pH7.6)
8M LiCl	6.25 (500mM)
0.5M EDTA	0.20 (1mM)
1% NP-40	1.00
0.7% Na-Deoxycholate	0.70g

Elution Buffer (50ml)	Volume (ml)
1M $NaHCO_3$	5.00 (100mM)
20% SDS	1.25 (0.5%)

This should be made up fresh each time.

EDTA-free protease inhibitor (Roche, cat number 04693132001) was added to all solutions just before use.

2.4.1. – DNA Extraction from ES cells

TE Buffer (pH8.0)	Volume (ml)
1M Tris HCl (pH8.0)	1
0.5M EDTA	0.2
H2O	To 100

Extraction Buffer	Volume (ml)
1M Tris HCl (pH8.0)	200µl (10mM)
0.5M EDTA	4 (0.1M)
RNase A (10µg/ml stock)	40µl (20µg/ml)
10% SDS	1 (0.5%)
H2O	To 20

2.4.3. - Southern Blotting

Denaturation Buffer	Volume (ml)
NaCl	58.44g (1M)
NaOH	20g (0.5M)
H2O	To 1L

Neutralisation Buffer	Volume (ml)
NaCl	87.66g (1.5M)
Tris	6.057g (50mM)
0.5M EDTA	2ml (1mM)
H_2O	To 1L

20x SSC	Volume (ml)
NaCl	175.3g
Trisodium citrate	88.2g
H_2O	To 800mls, adjust to pH7 (using 1M HCl)
H_2O	To 1L

Salmon sperm DNA

- Add 25ml H_20 to 250mg DNA to give 10mg/ml
- Pass through G18 needle several times to shear

Church Buffer	Volume (ml)
1M PB	15
0.5M EDTA	60µl
SDS	2.1g
BSA	0.3g
H_20	To 30

1M PB	Volume (ml)
Na_2HP0_4 pH8.0	67g
Phosphoric acid	2 (in fume hood)
H_20	To 500

Wash Solution 1	Volume (ml)
BSA (coarse)	2.5g
SDS	25g
1M PB	20
0.5M EDTA	1
H_20	To 500

Wash Solution 2	Volume (ml)
SDS	10g
01M PB	40
0.5M EDTA	2
H_20	To 1L

7.2. - Primers and Probes

2.3.5. - Quantitative Real-Time PCR

Custom Plus TaqMan assay

	Based on primers from
GAPDH	Kernohan, 2010
Gtlk_GD3	Kernohan, 2010
H19	Kernohan, 2010

Chen 2	Sequence
Forward Primer	CCAGGAATAAGCTTAGAGGCCTTTT
Reverse Primer	GGAATGTCTACCGGCCTACTC
FAM conjugated probe	CCGCTCTCAACCCTCC

H13e	Sequence
Forward Primer	GATCACTCAGCCCATTTCTGTCT
Reverse Primer	CTTTTCCTAGCCATTCCTCAGTCT
FAM conjugated probe	CCTGTTTGCAGTAGTCCAC

Using the Roche Universal Probe library

CpG negative region	Sequence
Forward Primer	CTCTGGTCCAGGGATTTGAA
Reverse Primer	AAAGCAACACACATCCACCA
Probe Number	7

MNT	Sequence
Forward Primer	CTGGGTCGCACGTCTAGC
Reverse Primer	AGCAGTCCGGGTAACCAAC
Probe Number	102

Nap1l5	Sequence
Forward Primer	TGGGCAAGCTCTCCATAAAG
Reverse Primer	CAGCTGAGCCGAGCAGTAG
Probe Number	32

2.4.3. - Southern Blotting

Fam13c	Sequence
Forward Primer	GAGACGACCCCACACATCAT
Reverse Primer	CTCGTCCCTTCCAATTCTTG
Probe	GAGACGACCCCACACATCATTCAGGCCCCATTTAATAA ACCCCCTTACGAGAGAGACCTGGGACCTCCAGTCAACC TGCGGAGTTCTAACTTCTCTGGTTACCAGATTATTGGAA ACAAAACCACCCTAATCCACTTCTCAGATCCAGAGCAC AGGTGAGTGTGCTAATCCACGCCAGGACTGCCATGTGG CTTGGACGTATTGAGCCATGCAGACGACTTGGCTCCCT GCCTCTGGGACCAATTCTCTTCTTCTTTTGTCCCTTTCTG GGATGGCAGTGTACTGTTACTGCTAAGGACAATCACTG TAGGCTCATACCGTTCTTTTCTTCCTTCCTTTGTTCTGTA GGTCCATATTTAAGATACAGGGGAGTTGGTCCAAATTT CTGGAAGTTTCATGACCAAGAATTGGAAGGGACGAG

Fam13c neo	Sequence
Forward Primer	CAACAGACAATCGGCTGCTCTG
Reverse Primer	GATAGAAGGCGATGCGCTGCG
Probe	CAACAGACAATCGGCTGCTCTGATGCCGCCGTGTTCCG GCTGTCAGCGCAGGGGCGCCCGGTTCTTTTTGTCAAGA CCGACCTGTCCGGTGCCCTGAATGAACTCCAGGACGAG GCAGCGCGGCTATCGTGGCTGGCCACGACGGGCGTTCC TTGCGCAGCTGTGCTCGACGTTGTCACTGAAGCGGGAA GGGACTGGCTGCTATTGGGCGAAGTGCCGGGGCAGGA TCTCCTGTCATCTCACCTTGCTCCTGCCGAGAAAGTATC CATCATGGCTGATGCAATGCGGCGGCTGCATACGCTTG ATCCGGCTACCTGCCCATTCGACCACCAAGCGAAACAT CGCATCGAGCGAGCACGTACTCGGATGGAAGCCGGTCT TGTCGATCAGGATGATCTGGACGAAGAGCATCAGGGG CTCGCGCCAGCCGAACTGTTCGCCAGGCTCAAGGCGCG TATGCCCGACGGCGAGGATCTCGTCGTGACTCATGGCG ATGCCTGCTTGCCGAATATCATGGTGGAAAATGGCCGC TTTTCTGGATTCATCGACTGTGGCCGGCTGGGTGTGGC GGACCGCTATCAGGACATAGCGTTGGCTACCCGTGATA TTGCTGAAGAGCTTGGCGGCGAATGGGCTGACCGCTTC CTCGTGCTTTACGGTATCGCCGCTCCCGATTCGCAGCGC ATCGCCTTCTATC

2.4.4. – Long-Range PCR

Primer Name	Sequence
H13 Screen F1	TTGAGAGGCTGAGGCGTGAGG
H13 screen neo R1	CGGGACTATGGTTGCTGACT
Conseq F10	CCGCTTTTCTGGATTCATCG
H13 Screen R1	CCCTCTGATCCTTGCCGTCTC

2.4.9 – Sequencing

Primer name	Sequence
Conseq F3	CCCGTCAAAATAGTGAGATGCC
Conseq F4	CTTCTACTCCTCCCCTAGTCAG
Conseq F5	GCAGCAAGAAGCCACGGAAG
Conseq F7	CGTGTTTGCCTGGGCTTTGG
Conseq F8	CCCCAGCAGGCAGAAGTATG
Conseq F9	GGTGCCCTGAATGAACTCCA
Conseq F10	CCGCTTTTCTGGATTCATCG
Conseq F11	TCTTCTGAGCGGGACTCTGG
H13 3’ seq F1	CTAGCTTCAGGTGGCAGGGC
H13 3’ seq F2	GCAGGATGGGACACAGCGAG
H13 3’ seq F3	GGTAAATGAGTTGAGGCTCAGG
H13 3’ seq F4	GTGTCTTAGTGGGATGGAAGC
H13 3’ seq F5	CATCATTGGTGTCAGGGACTC
H13 3’ seq F6	CAGGTCTGCACAGAGAAACCC
H13 3’ seq F7	CCACAGTTCTAGGGTTGACTCC
H13 3’ seq F8	CATGCAGGCAAAACACCAATG
H13 3’ seq R1	CAGCAAGGTAGATGGAGAGCG
H13 3’ seq R2	GGGGACCTGGATTCGATCTC
H13 3’ seq R3	CTCTCATCCTGCTGCTTCCGC
H13 3’ seq R4	CACTACATCACGCTGGGATTC
H13 3’ seq R5	TACTGCATCTAGAAGCCTAGGGG
H13 3’ seq R6	TGGTAGTAGTGGTGGTAGTGTCC
H13 3’ seq R7	CATGATGTCCAGCTCTGAGGG
H13 3’ seq R8	CCTCCTTCAGGCGCTGACAG
H13 3’ seq R9	CCAGGCACTGTGCTGAGGGCA
H13 5’ seq F1	CCACTGCCTGCCTCAGACCAG
H13 5’ seq F2	CTCCAGTTCACAGATTCACACC
H13 5’ seq F3	GAGGCAGGTCTGAGCTGTAGAG
H13 5’ seq F4	CTCTATGCCTTGTACCCGTCT
H13 5’ seq F5	GGCTATCACTAATGTAACCCCC
H13 5’ seq F6	GCAACAGAGACTTCATAGCAGGC
H13 5’ seq F7	CAGTGCTGCCCATGTCCCCT
H13 5’ seq F8	CGATGCCTTGGACTAAGTGG
H13 5’ seq R1	GGAAGCAATCAGCACAGTGCC
H13 5’ seq R2	GCAACAGAGACTTCATAGCAGGC
H13 5’ seq R3	GCAACAGAGACTTCATAGCAGGC
H13 5’ seq R4	GATGTTTCCAGACCTACCTTG
H13 5’ seq R5	GGCAGAGATGGGACTTGAACC
H13 5’ seq R6	CCTGAGCAACTGAGTGAAACTC
H13 5’ seq R7	CCTACTCAAGTGGTCCACAGC

2.4.10.4.1. – PCR

Primer Name	Sequence
H13i4/x5 F1	CAGCGATCGCACATCATGAAGACAATACCTC
H13i4/x5 R	CAGGCCGGCCCAGGCACACCAGGTCCTT
SH_modi3_F	CAGCGGCCGCCAGTTGCAGGCAGGTAGGAGAC
SH_modi3_R1	CAACGCGTCACCTCCAGCTTTTGATCC

2.4.10.4.2. – Long-Range PCR

Primer Name	Sequence
SH_i3i4_F1	CAACGCGTCACCAGCCTGAGCAACTGAGTGAAAC
SH_i3i4_R	CACTCGAGCATGGTTGCTATCCCAGACATCC
SH_i3_F1	CAGATATCCACTATGGGCTTTGGAGTAAATCTAGC
SH_i3_R	GTGCGGCCGCGTGCTGTGAGCTGTGGTCTGGTC
SH_x3-i3_F	CAGTCGACCAAGGCATCTCTCGACAGCCT
SH_i3_R	GTGCGGCCGCGTGCTGTGAGCTGTGGTCTGGTC

2.4.10.14. – Sequencing

Primer Name	Sequence
eGFPf	GAGCAAAGACCCCAACGAGA
eGFPf1	GAACACCCCCATCGGCGACG
eGFPr	AACTTCAGGGTCAGCTTGCC
eGFPr1	GTCACGAGGGTGGGCCAGGG
SH_i3i4SEQ_F	ACGTCCATAAAGCCGTTCAG
SH_i3i4SEQ_R	TGTCTGCACTGATTTCTGTATGTG
SH_i4x5SEQ_F	CTCTGCCCAGAGATCATCAA
SH_i4x5SEQ_R	CCCACTCTGGATTCTTCCAA
SH_mCherrySEQ_F	CACCATCGTGGAACAGTACG
SH_mCherrySEQ_R	CGCCCTCGATCTCGAACT
SH_modH13i3_F	CTTGCCTTGATCCCAACTGT
SH_modH13i3_R	TTTTGGTAACTATTAGGTGAA
SH_polyASEQ_F	CCCAGGTCTCCCAAAATACA
SH_polyASEQ_R	GGCCATGGGATGAGTTTTTA
SH_pTRE3G_f	CGAGGCCCTTTCGTCTTCAA
SH_pTRE3G_R	GGGAGTAAACTCTTCATACGTTCTC
SH_pTRE3GSEQ_F	CTGGAGCAATTCCACAACAC
SH_UCOE_F4	GGAAAAGACATTGGTCCCCT
SH_UCOE_R5	CAGTTCTCACTACAGCGCCA

2.4.11.5. – Quantitative Real-time PCR

Using the Roche Universal Probe library

i3	Sequence
Forward Primer	aaaatgcccgctaaagactg
Reverse Primer	gagggatttgggtgtggtc
Probe Number	49

x2	Sequence
Forward Primer	acatgccagaaaccatcacc
Reverse Primer	aagaggtagagccccaggag
Probe Number	60

x3	Sequence
Forward Primer	tgggaacatggtcttttgtg
Reverse Primer	gggaagattcccttggattt
Probe Number	21

x5eGFP	Sequence
Forward Primer	tcgtgaccaccctgacctac
Reverse Primer	aagtcgtgctgcttcatgtg
Probe Number	41

mCherry	Sequence
Forward Primer	gaagggcgagatcaagca
Reverse Primer	ttgacctcagcgtcgtagtg
Probe Number	41

7.3. - Restriction Enzymes

All restriction enzymes, BSA and buffers used were purchased from New England Biolabs®.

Restriction Enzyme	NEB catalogue number	Method used in
*Age*I	R0552S	2.4.10.6.
*Ase*I	R0526S	2.4.10.6.
*Asi*SI	R0630S	2.4.10.6.
*Bam*HI-HF	R3136S	2.4.10.6.
*Bgl*II	R0144S	2.4.10.6.
*Eco*RI	R3101S	2.4.10.6.
*Eco*RV	R0195S	2.4.10.6.
*Fse*I	R0588S	2.4.10.6.
*Hind*III	R0104S	2.4.2.
*Kpn*I	R0142S	2.4.10.6.
*Mlu*I	R0198S	2.4.10.6.
*Not*I	R0189S	2.4.10.6.
*Pci*I	R0655S	2.4.10.6.
*Pac*I	R0547S	2.4.10.6.
*Sac*I	R0156S	2.4.10.6.
*Sal*I	R0138S	2.4.10.6.
*Sal*I-HF	R3138S	2.4.10.6.
*Spe*I	R0133S	2.4.2.
*Stu*I	R0187S	2.4.2.
*Xba*I	R0145S	2.4.2.
*Xma*I	R0180S	2.4.10.6.
*Xho*I	R0146S	2.4.10.6.

Publications

Prickett, A. R., Barkas, N., McCole, R. B., Hughes, S., Amante, S. M., Schulz, R., Oakey, R. J. (2013) Genome-wide and parental allele-specific analysis of CTCF and cohesin DNA binding in mouse brain reveals a tissue-specific binding pattern and an association with imprinted differentially methylated regions. Genome Research, 10, 1624-35.